Mobility and Spatial-Temporal Traffic Prediction in Wireless Networks Using Markov Renewal Theory

by

Haitham M. Abu Ghazaleh

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Computer Engineering

University of Manitoba

Winnipeg, Manitoba, Canada

TABLE OF CONTENTS

ABSTRACT

An understanding of network traffic behavior is essential in the evolution of today's wireless networks, and thus leads to a more efficient planning and management of the network's scarce bandwidth resources. Prior reservation of radio resources at the future locations of a user's mobile travel path can assist with optimizing the allocation of the network's limited resources. Such actions are intended to support the network with sustaining a desirable Quality-of-Service (QoS) level. To help ensure the availability of the network services to its users at anywhere and anytime, there is the need to predict when and where a user will demand any network usage. In this thesis, the mobility behavior of the wireless users are modeled as a Markov renewal process for predicting the likelihoods of the next-cell transition. The model also includes anticipating the duration between the transitions for an arbitrary user in a wireless network. The proposed prediction technique is further extended to compute the likelihoods of a user being in a particular state after N transitions. This technique can also be applied for estimating the future spatial-temporal traffic load and activity at each location in a network's coverage area. The proposed prediction method is evaluated using some real traffic data to illustrate how it can

lead to a significant improvement over some of the conventional methods. The work considers both the cases of mobile users with homogeneous applications (e.g. voice calls) and data connectivity with varying data loads being transferred between the different locations.

ACKNOWLEDGMENTS

I am heartily thankful to my research adviser Dr. Attahiru Sule Alfa for his endless support, guidance, and encouragement. I owe my deepest gratitude for all his assistance and teachings. This project would have not been possible without him.

I further extend my gratitude for all the encouragement, continuous feedback, insightful comments and suggestions that have been shared with me by the thesis committee.

I am also grateful for all the assistance given to me by the academic professors, staff, and colleagues in the Electrical and Computer Engineering department at the University of Manitoba. I am especially thankful for the added support of the staff and colleagues at TRLabs-Winnipeg who have also provided me with a valuable experience and the opportunity to expand on my research throughout my PhD program.

I would also like to thank CRAWDAD (Community Resource for Archiving Wireless Data at Dartmouth) for their support and for sharing their data on wireless user mobility and traffic with the rest of the research community. Their shared resources have been instrumental in this research work.

Above all, I am utmost thankful to be blessed with the boundless love and support of my parents, Mahmoud and Makarem, and my sister, Haya, who have done the impossible to help get me to where I am today. I dedicate my achievements in life to them.

This research work is supported in part by a Grant from NSERC (Natural Sciences and Engineering Research Council of Canada) to A. S. Alfa.

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background and Motivation

The rapid growth of mobile networking and the diversity in network applications have prompted the need for future generation wireless networks to support a range of Quality-of-Service (QoS) levels. Wireless network activity, as well as the number of users, are expected to continually increase with the gradual development of diverse wireless applications that demand high bandwidth. One of the studies, the "COIN" project (Dynamics of COmpetition and INnovation in the converging Internet and mobile networks) [1,2] conducted in Helsinki-Finland included an extensive analysis of some data received from local mobile managers, specifically during the period of 2005 and 2007. Amongst their many results, they were able to show how the number of users with mobile connectivity have increased over the years. Their results have also shown a parallel growth in users desiring mobile terminals with data transmission capabilities. This had led to the conclusion that more users are favoring wireless connectivity and preferring the usage of mobile devices due to the freedom it provides in terms of supporting various services while roaming between multiple locations. While such services may have been a luxury in the past, the need

for mobile connectivity continues to readily blend in with many of today's cultures and generations.

With the rising trend of enjoying wireless access anywhere and anytime, mobile users are becoming more concerned with the QoS levels that can be supported by the network. Hence, future wireless networks are expected to both improve and safeguard these QoS agreements despite the users' movements and the network's traffic. One method for ensuring the availability of the network services anywhere and anytime to a user is to predict at anytime, and within reason, where a user will likely demand the network usage. In wireless and/or mobile networks, the available bandwidth within the coverage area is the main resource under contention and is of primary concern to the network managers due to how scarce it is. The availability of such resources will somehow need to be guaranteed to avoid any abrupt disruptions in a user's ongoing and active connection while being mobile. However, such assurances do come at a cost to the network and/or user with the aim being at trying to minimize such costs without placing any burden on the overall performance.

Wireless networks are generally divided into distinct geographical units known as *cells*. Each cell provides a wireless coverage that is administered by a single access point or *base station*. The wireless bandwidth that is available at each of the coverage locations can be shared and administered under various schemes such as Time-Division Multiple Access (TDMA), Frequency-Division Multiple Access

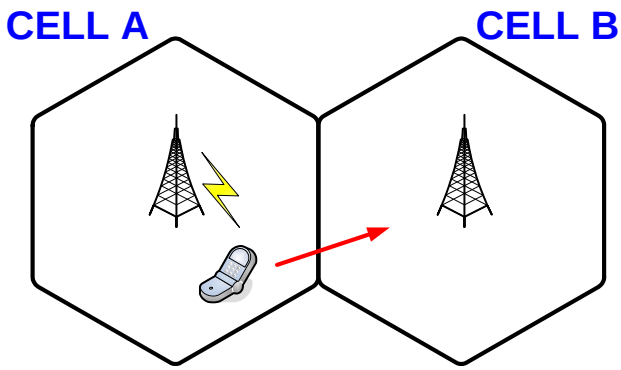


Fig. 1.1: A mobile unit generating a handoff request into Cell B.

(FDMA), Code-Division Multiple Access (CDMA), and many others [3, 4], as well as a hybrid of these schemes. However, each of these schemes is limited in how much of the available resources they can allocate to users within a coverage area, thus restricting the number of users that can simultaneously connect with an access point to some finite capacity.

Occasionally, mobile users may need to have their ongoing connections transferred between different base stations (or cells) in order to maintain their active sessions; a process known as *handoff* [5]. For example, Figure 1.1 shows a mobile user that is engaged in an active session with the network and utilizing the wireless resources of Cell A. If the same user wishes to maintain its connectivity with the network when moving to Cell B, then the network should seamlessly transfer the user's ongoing active session to the new cell. Successful handoffs are only possible if sufficient resources can be granted to the ongoing session by the new network access point. Otherwise, the session will be prematurely terminated, or *dropped*, as a result of insufficient resources available at the new cell [6]. A new call/connection that

attempts a resource request at a coverage location that has reached its capacity in terms of the resources that it has allocated to others will be *blocked* from service and the user can retry its request at a later time. While both *blocked* and *dropped* connections have the same effect of denying a user from connecting with a particular access point, it is generally perceived that the latter is less desirable from a user's perspective.

Other than increasing the amount of bandwidth available at each of the coverage areas, which is not commonly feasible, the overall capacity of the network can be increased by reducing the actual coverage area size of some/all access points while the bandwidth remains the same. This has the effect of increasing the capacity per unit area since the coverage area has fewer users to consider. Even though this may seem as a reasonable approach for stretching the amount of network resources available to the users, it does however lead to more chances of a mobile user initiating a handoff request. Increasing the frequency of such requests could lead to the undesirable possibility of having the connection dropped [3, 4]. Regardless of the approach taken to improve the capacity, there still remains the need for the network to efficiently administer its resources at various locations. The ultimate purpose of efficient resource management scheme is to maintain an upper level of quality and performance for both the users and the network.

The network's access points are responsible for efficiently managing their limited wireless resources while adhering to the following requirements: 1) maximum uti-

lization of their resources which in turn maximizes the network manager's revenue, and 2) optimum resource allocation in order to minimize both the new call blocking and dropping of handoff requests. The latter facilitates with both improving the user's experience with the network and minimizing the lost revenue of the network managers. This is based on the assumption that blocked and dropped calls tend to discourage users from associating with the network and thus having a negative impact on revenue. The best tradeoff can be achieved if the user's mobility behavior was known to the network, thus allowing for sufficient resources to be reserved at the right places and times. In other words, if the time and place of a mobile user's next cell transition is known, then the network can proceed with examining the possibility of reserving the necessary resources at the new cell location. This is accomplished in an attempt to prevent the user's active connection from being dropped at the new location. Such actions can assist with maintaining a certain level of QoS for the mobile user throughout the lifetime of its connection with the network.

Network managers are usually concerned with the connection dropping rates and the new connection blocking rates, with the aim of minimizing both. However, there tends to be a much higher interest in maintaining a dropping rate below some certain level. Various call admission control schemes such as "Guard Channels" assignment (fraction of total channels reserved for a certain class of connections) and prioritizing of handoff requests [6,7] have been shown to reduce the connection dropping rates at

the cost of new connection blocking rates. In such schemes, the resources set aside for higher priority connections (i.e., handoff requests) are typically assigned for long periods of time and not reserved for any specific user. Some of these reservations can be unused for significant amounts of time when it could have been better utilized by other lower priority connections. Therefore, a better approach would be to only reserve these resources when potentially needed by a specific user, which would tend to lessen the excessive reservation of resources.

For user-specific resource reservations, one way of minimizing the dropping of handoff requests is to reserve enough resources at *all* neighboring cells for the handoff request. However, such an approach leads to a wastage of resources along with potentially increasing the blocking rate for new connection requests. A better alternative would be to limit the reservation of these resources to the regions of the network where a user is *likely* to visit. Therefore, an efficient and accurate prediction of a mobile user's transitions is vital since misplaced reservation of network resources will not only fail to uphold the desired QoS, but also likely degrade the performance of the overall network. In fact, it has long been accepted that call admission control which is aided by mobility prediction techniques can significantly enhance the network's performance [8]. For example, it has been shown in [9] and [10] how the QoS in wireless networks can be considerably improved when applying mobility prediction algorithms on the network's call admission control. The further ahead the prediction scheme tries to pre-allocate the resources, the more likely a network can

honor its QoS agreement for the lifetime of a user's session.

However, the task of predicting how far a network should proceed with reserving the resources can be quite challenging, especially when the connection lifetime is unknown. Moreover, the efficiency of any prediction scheme greatly depends on how likely the resource reservations are made at the right place and time. The type of information to be used for making a prediction is also very crucial on determining the appropriateness of the prediction scheme.

Other usages for mobility prediction have been explored by a few researchers such as Gossa et al. [11] who looked at applying it to data replication management in distributed mobile computing scenarios. In this thesis, while the proposed mobility predictor could be applied for other contexts, it will mainly be presented for the purpose of enhancing the allocation of resources for mobile users with handover requests [12].

1.2 Previous Work

Over the recent years, there has been a considerable amount of work done on developing mobility prediction and network traffic estimation techniques. This is due to the increasing need in efficiently managing the network's resources for a continually growing number of users. Prediction schemes have been proposed for both infrastructure-based and infrastructure-less wireless networks. Examples of prediction schemes for infrastructure-less wireless networks, such as Ad Hoc networks, have been proposed by Pathirana et al. [13], Su et al. [14], and Chellappa et al. [15, 16]. They are quite complex to model due to the dynamic topology that may change rapidly and unpredictably. In this thesis, it was chosen to focus mainly on developing mobility models for wireless networks with static infrastructures.

Various mobility prediction schemes have been proposed with many of them relying on the availability of prior information on the user's mobility behavior. The models presented in [17] and [18] are examples of prediction schemes that require no knowledge of the user's mobility history. In [17], the constant tracking of the relative distances between the mobile user and the neighboring access points is proposed, with the potential access point predicted as the one where the distance to the user falls below a certain threshold. This avoids having to keep a record of the mobility pattern of the users. In [18], it was proposed to additionally monitor a user's preferred movement direction for making any predictions. Jayasuriya and Asenstorfer [19] also looked at tracking a user's geographical location, speed and direction

of movement relative to the neighboring access points and developed a model for predicting a mobile user's transitions into the neighboring locations. However, the authors also mentioned the difficulty with practically obtaining an accurate measure of a user's location and speed using the received signal strengths from the access points. While the continuous tracking of mobile users may lead to better predictions in terms of movement, such schemes will likely suffer from the large overhead accrued due to the constant monitoring. Moreover, erratic user movements could also downgrade the performance of such schemes.

Many of the prediction techniques rely heavily on historical data that include information on aggregate mobility and handoff history at each location. Examples of such models include those given in [20–24]. Chan et al. in [20] and [21] have proposed various prediction schemes based on some mobility history that contain records of a user's next move, direction of travel, as well as other information. These proposed schemes have proven to be the basis of the many prediction techniques that followed, but are only limited to predicting *where* (without *when*) a user is likely to move. To combat the instances where their prediction schemes may perform quite poorly due to abrupt variations and randomness in user behavior, the authors further proposed to consider a fraction of all the neighboring cells as part of the prediction result as opposed to a single cell and based on a pre-defined "Prediction Confidence Ratio (PCR)". A similar approach was taken by Erbas et al. [25] and they proposed that it was sufficient to only consider up to the most likely 3 of the 6 neighboring

cells as the prediction result. In [26], Capka and Boutaba proposed a mobility predictor using neural networks which is capable of both learning and predicting future transition patterns of mobile users. In their results, the authors have noted that better neural network predictors would ultimately require a substantially large neural net for capturing the complex nature of user mobility.

Prasad Agrawal in [27] proposed a second order Markov chain predictor while Song et al. [28] discussed the advantages of an order-k Markov mobility predictor and have tested it with actual mobility data. Such predictors were shown to perform quite well for lower orders of k, but did not perform as well with predicting *when* a movement would occur. Furthermore, an immense amount of mobility history is needed for generating higher order-k Markov mobility predictors. An order-k Markov mobility predictor *with fallback* was proposed by Sun and Blough [29], where the predictor falls back to a lower order of k if a certain order-k predictor was unsuccessful. Their work was limited to making location predictions alone. In [30], Yu and Leung (and similarly with the work done by Bhattacharya et al. in [31]) proposed applying data compression methods (namely Ziv-Lempel) on the available mobility history data and developed their prediction scheme accordingly. They based their proposal on the rationale that *"good data compressors should also be good predictors"* [30], as well as on the basis that *"the repetitive nature of mobility patterns suggests the stationarity of a sequence of events generated by an m^{th} order Markov source and based on the previous m events"* [30]. However, the generation

of the Ziv-Lempel tree from the mobility history can be quite complex. This idea was also considered by Song et al. in [32] and further examined the efficiency and accuracy of such methods with actual mobility data.

A few models considered utilizing both the mobility historical data as well as the current conditions in the network. One example is the two-stage mobility prediction developed by Park et al. in [33] which considered both the location transitions modeled as a Markov process and the user's geographical movement. Another example is the model proposed in [34] by Akyildiz and Wang which considered both the velocity and direction of mobile users as well as historical data for predicting the future locations. This work was preceded by the proposed "Shadow Cluster Concept" given in [35] which estimates the fading likelihood levels of moving to the neighboring cells and further, much like a shadow. Their scheme could predict both future locations and service requirements, but does not give any indication on when these changes in location are likely to occur. In [36], a hierarchical location prediction model was proposed which employs mobility history for predicting inter-cell movements while considering the mobile user's speed and movement direction within the cell. Their model is limited to location predictions alone. Another hierarchical model developed by Wang et al. in [37] proposes to model both the macro-location and micro-location mobility behavior of a user along with incorporating the movement velocity and direction into the model. The difficulty with such hierarchical models is with distinguishing between what constitutes as macro-mobility and micro-mobility.

Incorporating road topology information into the mobility prediction schemes, as proposed by Soh and Kim in [6, 38] and Lee and Hsueh [39], can improve on the prediction results by affirming or even eliminating certain paths. This approach can be of particular benefit for modeling vehicular mobility. Samaan and Karmouch [40] considered further utilizing a geographic map with identifiable landmark objects (e.g., restaurants and malls) into the user-mobility predictions and for better characterizing the user's mobility behavior. This was based on the assumption that the majority of movement patterns are related to user-activity and purpose-of-mobility [41] (i.e., users move with a particular destination in mind). However accurate their prediction results may be, such schemes require a vast amount of information to be collected and processed and may not perform very well with temporary changes of the surrounding infrastructure. In [42], each user's daily itinerary patterns (including both location and time) were proposed to be incorporated into their location area predictions. Such information could be quite challenging to acquire, especially when dealing with user privacy issues, and may not be suitable for public wireless networks. Another example is the model proposed by Ghosh et al. in [43] which considers the "mobility profile" of the users in terms of the regular places visited (e.g., office, home, shopping centers) and exploiting this information for better predicting the future location of the user in the network. While this approach may be beneficial for estimating the mobility patterns on a per user basis, the proposal given in this thesis is mainly focused on developing mobility models

for an aggregate number of users.

Other works include those that proposed elaborate user mobility models that depict near realistic behaviors. With such models, network providers can gain some understanding on how to manage their resources for optimum performance. An example of such a model is given by Stepanov et al. in [44], where the authors model the mobility of users in outdoor scenarios which takes into account various factors such as environmental constraints, user travel decisions, and the associated activities at the many locations. For indoor scenarios, Lessmann and Lutters in [45] developed a user behavior model that considers environmental constraints, activity variations and movement patterns that can either be based on some schedule or according to some stochastic process. Most of these models specifically address the issue of user mobility in ad hoc networks where it is crucial to have a good understanding of the exact movement behaviors. However, the focus of this thesis is on networks with centralized control, such as cellular networks. In such cases, it is sufficient to examine user mobility from the network's perspective which is not the same as looking at the exact user mobility behavior for reasons given in the next section.

An ample amount of research has been devoted to thoroughly analyzing vast volumes of real wireless network measurements for extracting and attempting to characterize the behavior of users in such networks and for the purpose of modeling their mobility patterns. They include the works done by Ghosh et al. [43],

Papadopouli et al. [46], Chinchilla et al. [47], Kim et al. [48], Boc et al. [49], Yoon et al. [50], and the list is certainly not limited to them. In addition to analyzing the behavior of the web-traffic requests, Chinchilla et al. [47] used the WLAN traffic data available to them and modeled the user location transitions as a simple Markov chain for predicting the next location transitions alone. In [46], Papadopouli et al. employed graph theory to model the roaming behavior of the users using data collected from real Wireless Local-Area-Network (WLAN) traffic. Their data were used to analyze the degree of connectivity of the node pairs on the graph, where each node represents the access points in the network. The degree of connectivity corresponds to the likelihood of making a transition from one access point to another. Their interest was mainly in studying the impact of the changes in the graph topology (i.e. wireless infrastructure) on the degree of connectivity. Their results could also be used to estimate the spatial transitions of the users but not temporal. Kim et al. [48] proposed analyzing the WLAN traffic traces for extracting a user's mobility tracks which describes the pathways taken by a user that is roaming between various access points. These tracks can then be used to further extract certain characteristics of the user's mobility behavior which can be used to generate future mobility tracks that closely resemble the estimated behavior of the user. The authors were mainly focused on generating a mobility model for simulation purposes and could further be applied for estimating a user's future transitions. However, their model does not consider any temporal influences.

Network managers have also shown an interest in analyzing spatial traffic behaviors of their mobile users for the purpose of efficiently managing the limited resources. However, the research on traffic estimation is usually independent from those on mobility prediction. In [51], Adas reviewed many of the traditional traffic estimation techniques and mainly focused on autoregressive models which typically require substantial amount of computations. The work done by Suzuki in [52] is one simple example where autoregressive models were used to study the average daily vehicle traffic. Another example includes the work done by Bermudez et al. in [53] that proposed to model the periodically-varying population distributions at distinct locations based on the competition principles of biological species. However, these works ignore the temporal influences on the spatial traffic and focus mainly on long-term behaviors. Borrel et al. [54] and Zhou et al. [55] had further looked at modeling the changes in the spatial behavior for groups of mobile users that tend to gather towards some attracting point. In addition, Chen et al. in [56] proposed a technique for generating a traffic volume forecasting model that considers estimating various details from a series of trends deduced from a temporal traffic volume data using the wavelet transform. Their results were subsequently used to train a neural network for predicting such trends. While their approach has been validated with real traffic data, their technique strongly relies on the suitable choice of numerous parameters that are required for optimal trend estimation.

Several researchers considered analyzing real traffic records for extrapolating

some trends and characteristics to be used for modeling the spatial behaviors of the users in the network. The analysis of billing records of a CDPD wireless network done by Andriantiatsaholiniaina and Trajkovic in [57] revealed some interesting trends on the spatial traffic behavior of the network's mobile users and offered some guidance on how to model such behaviors. Similar achievements have been made by Hutchins and Zegura [58], Tang and Baker [59], and Balachandran et al. [60] and focused on analyzing the traffic in WLANs. Almeida et al. [61] considered the analysis of GSM traffic and proposed a model for describing both the spatial and temporal network traffic distributions. Based on the results reported in [61], the authors concluded that the spatial traffic can be modeled using exponential/linear and piecewise/linear functions, while the temporal traffic was based on either the double-gaussian or trapezoidal functions. Most of the conclusions that have been derived by those researching the trends of some real traffic data have generally yielded a better understanding of specific behaviors. However, their conclusions are highly dependent on the type of network that they are investigating and may not be valid for others.

Ning et al. [62] proposed employing bilinear interpolation for deriving a continuous model of only the spatial traffic behavior in the network. They further proposed the construction of a visualized representation of their spatial analysis using pseudo color and contour maps. Tutschku and Tran-Gia [63] considered taking a map of the network's coverage area and characterizing the entire area using discrete demand

nodes. These nodes were then used to model the spatial network traffic demands at these discrete points for the purpose of optimizing the network coverage areas. Ashtiani and Salehi [64] proposed a stochastic model for capturing the steady-state spatial traffic distribution of mobile users with multiple classes of services. Much of the work found were focused on estimating only the spatial traffic fluctuations for analyzing long term behaviors, but tends to ignore the influence of temporal fluctuations which are needed for short term analysis. The benefits of having such information on the spatial-temporal traffic fluctuations have been explored by Hampel et al. in [65].

1.3 Objectives

The focus of this thesis is to show how to better model the mobility behavior of users in a wireless network as a semi-Markov process. The model is primarily applied for predicting the next-cell transitions, along with anticipating the duration between the transitions for an arbitrary user in a wireless network. We ensure that the parameters to be used in the model can be derived from the wireless data that can be readily obtained from traffic logs. In essence, simplicity together with effectiveness is one of the major goals of this thesis. To the best of my knowledge, the only work found applying semi-Markov processes for mobility prediction was in [66] (with a similar idea proposed by Song et al. which they denoted as MarkovCDF in [28]) for next-location predictions alone. The model proposed in this thesis goes further and can deal with the case of both single and multi-transitions.

Single transition prediction considers estimating the next event (e.g. movement into the neighboring cell) whereas the Nth transition prediction considers estimating the future event after N transitions. The results of N transition predictions can be employed for estimating the travel path of a mobile user during its session lifetime. This information can potentially be applied for ensuring some level of end-to-end QoS guarantees, which is similar to what the model proposed in [35] tries to achieve. This is based on the rationale that the further ahead the network can predict when and where the user will likely transition towards (after multiple transitions), the better the chances it has at avoiding any disruptions in the continuation of a user's

ongoing active session. Only those disruptions that are due to the lack of availability of the necessary resources while transitioning between numerous access points is considered in this thesis. The models proposed by Le Grand and Horlait in [67] and Benmammar and Krief in [68] are good examples which illustrate the need for N-transition prediction for the purpose of managing the network resources and to support end-to-end QoS. Another example is the model proposed by François and Leduc in [69] which also looks at N transition predictions but without any temporal considerations. Note that the different types of resources to be efficiently managed by the network include the wireless and/or wired bandwidth, as well as others that are specific to the type of network that is addressed in the model.

In addition to predicting the location transition behaviors of the wireless users, it will further be shown how to employ the proposed mobility model for predicting other network-related characteristics that would be of significant benefit to both the user's and network's performance. For example, the model can also be used to estimate the expected traffic load and activity at each location in a network's coverage area, both spatially and temporally. Throughout this thesis, some numerical examples are provided to show how the proposed mobility model can be applied and how to interpret the predictions that can be computed from the model. Moreover, and wherever possible, the proposed prediction scheme is examined and validated using real wireless traffic traces. It will also be shown how the model could be utilized by the network managers for improving on their resource management.

Much of the work covered in this thesis, as well as the related work by other researchers, focus mainly on the next-location prediction of its network users that run some form of homogeneous application such as voice-calls. However, with the increasing traffic volumes generated from data connectivity in networks such as 3G and beyond, there is also the growing need to understand how these traffic loads are distributed across the network both spatially and temporally. Hence, there is the need to predict the transfer of these loads from one location to another. In this thesis, the mobility behavior of users with data connectivity and varying traffic loads is also proposed and modeled as a Markov renewal process. Some examples are given to show how the results from the model can be applied for spatially and temporally predicting the transfer of such loads between the different locations in the network.

1.4 Outline

Chapter 2 discusses the details of how to formulate the mobility behavior of users in a wireless network and how to model it as a Markov renewal process. This chapter also discusses how to generate the prediction results from the proposed mobility model along with analyzing the accuracy of these predictions returned by our proposed model.

In Chapter 3, the model discussed in Chapter 2 is tested using actual mobility traffic traces that were collected from another independent project. The traffic traces were used to derive the required parameters for constructing our proposed model which was then used to assess the accuracy of our predictions. The results of the predictions using the proposed model were compared with those obtained from a number of conventional prediction schemes developed by other researchers. A brief description is also included on how the prediction results can be interpreted and utilized for network resource management purposes.

Chapter 4 focuses on how to apply the proposed mobility model and extend the predictions for making spatial and temporal forecasts after N transitions in the future, which could aid with end-to-end QoS assurances. It will further be shown in this chapter how these multi-transition predictions can be applied for network resource reservation purposes and for estimating the network's spatial-temporal traffic characteristics. The traffic traces analyzed in Chapter 3 were again used to illustrate how to interpret the results returned from these multi-transition predictions. The

result can further be used to compute end-location predictions. Network managers can utilize the knowledge of the end-location predictions for extrapolating the set of possible locations that are likely to be visited by a user during the lifetime of its active session. Hence, this information can be used to reserve the needed resources at the set of multiple locations to assist with end-to-end QoS assurances.

Before concluding and summarizing the work in Chapter 6, Chapter 5 presents the details of how to apply the idea of our proposed mobility model and extend it for the case of modeling the mobility behavior of users that are running data-driven applications. In such cases, the concern is more on the amount of data traffic that a user carries from one location to the next. Some numerical examples are also given to illustrate a few of the inferences that can be deduced from the results of this model.

2. MOBILITY PREDICTION MODEL

2.1 User Mobility Patterns

One common misconception is the idea that user mobility patterns are close to being random [91]. Some researchers have found that mobile nodes exhibit some degree of regularity in their mobility patterns [31], which can be exploited for the purpose of predicting the future travel path of mobile users. In fact, mobile users are believed to travel on some defined paths with a pre-determined destination in mind, in accordance with their lifestyle and common trips. This does not imply that it is impossible for mobility patterns to exhibit random excursions, but instead suggests that such random patterns are rare. Moreover, mobility patterns tend to be significantly influenced by some geographic limitations, e.g. pathways and corridors.

In this thesis, an entire network space is assumed to be divided into zones and zone-IDs are used to specify the user's locations, e.g. the cell-ID in a cellular network that is currently serving the user. An example of such a representation is given in Figure 2.1 which shows a typical cell structure in mobile cellular networks, while other structures do also exist. Alternatively, a user's location can be identified by his/her geographic n-dimensional coordinates. Note that there exists a direct

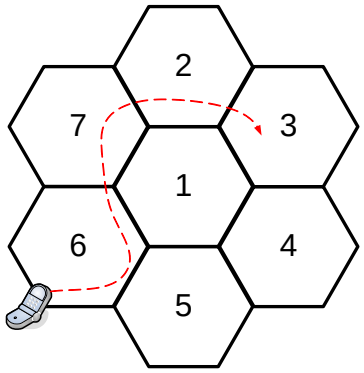


Fig. 2.1: 7-Cell network, with cell-IDs used to describe the terminal's mobility path.

mapping from the user's geographic coordinates to the zone-ID in which the user is located, while the reverse is not true. Since we are dealing with mobility predictions for assisting with the management of the network's resources, it is assumed that the tracking of the user's location via cell-IDs is sufficient from the network's perspective. Furthermore, the movement of a mobile user through the network can be described by the successive list of cell-IDs. This list represents the sequence of access points that were associated with the user's terminal throughout the lifetime of the active connection. For example, Figure 2.1 shows the user's mobility path to be $6 \rightarrow 7 \rightarrow 2 \rightarrow 3$. Each cell can be further divided into a number of sectors with distinct sector-IDs which are used to additionally describe the intra-cell movement patterns, as proposed by Kwon et al. in [23] with an example given in Figure 2.2.

The previous method of characterizing a mobile user's path and location is most suitable for networks with fixed infrastructures where the coverage areas that are supported by the various access points are static. However, other network types (such as Ad Hoc) have dynamically changing topologies which makes the process

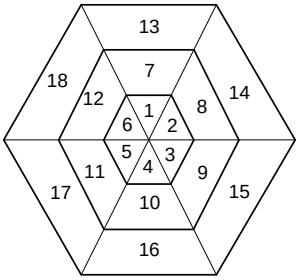


Fig. 2.2: A Zone-based sectoring of a hexagonal cell.

of tracking the user's mobility behavior quite challenging. For such networks, the network region can be segmented into smaller discrete (and possibly uniform) regions. The result is a grid of locations that could be used to characterize the zone at which the mobile user is located, with each zone having its own distinct identification number. An example is given in Figure 2.3. Furthermore, in such infrastructure-less networks, the coverage areas may vary quite considerably which could influence the mobility and traffic behavior of its users. The model proposed in this thesis mainly considers networks with fixed infrastructures in order to avoid such complexities. However, the methods discussed in this work may still be of some use for predicting the mobility patterns of users in infrastructure-less networks.

A user's mobility pattern from the network's perspective is determined by the user's *terminal* (e.g. mobile phone) mobility pattern. The regularities found in the movement patterns of users are not necessarily similar to those in the connection trace of a mobile terminal. The transfer of a user's connection to a neighboring access point could be for reasons other than the user moving out of the current cell such as signal fluctuations, congested cells, or constraints in the surroundings.

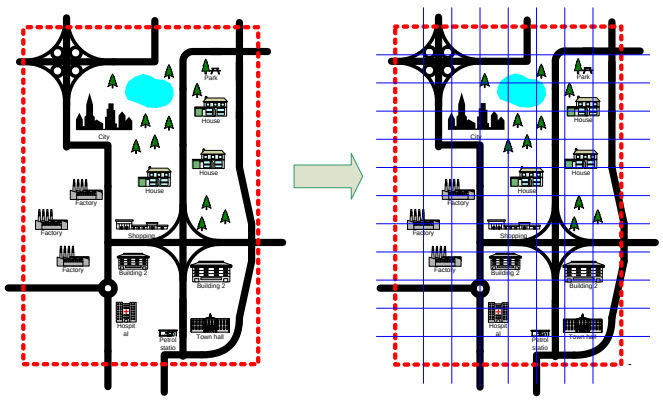


Fig. 2.3: The partitioning of a geographical region into smaller zones.

The users' mobility history patterns can be periodically recorded using the cell-ID representations. Let $P_{i,j}$ be defined as the probability of making a transition from cell i to cell j, and $\tau_{i,j}$ as the time spent in cell i before making a transition to the neighboring cell j. For each cell that is accumulating a mobility pattern profile, the number of handoffs made to a neighboring cell, as well as the time spent in the current cell before the transition occurs, can be recorded. This allows for the computation of the cell-transition probabilities $P_{i,j}$, as well as $\tau_{i,j}$. In addition, the distribution of the cell sojourn times at each location can be extrapolated from the set of recorded sojourn times. We assume that the network at each of its locations keeps a record of each session's sojourn time in the cell and the cell-ID of the next cell transition.

The mobility history can either be recorded for each user, or collectively for all users, into a single history profile per location. The latter method is more suitable for situations where all users will generally exhibit similar behavior at a given access

point, and are also not significantly influenced by erratic behaviors from a single user. Even though different groups of users have different mobility patterns, it can be difficult to address every type of group behavior in a single mobility model. However, it is assumed that an aggregated mobility history of users account for the many different types of behaviors, as argued by Capka and Boutaba in [26]. A large enough history of traffic traces could help with ensuring the capturing of the many different group mobility behaviors. Alternatively, each group of users with a shared mobility behavior can be modeled as a separate class of users with its own mobility model. In this thesis, and for simplicity, it was chosen to base our proposed mobility model on the aggregated behavior of the mobile users.

It has been shown in [21] that the accuracy of the prediction for the next-cell transition can be improved by additionally considering the prior location of the user immediately before making the transition into the current location. Define $P_{h,i,j}$ as the probability of making a transition from the current cell i to cell j, given that the user was in cell h prior to being in cell i. Let $\tau_{h,i,j}$ be the time spent in cell i prior to making a transition that is given by the probability $P_{h,i,j}$. Hence, both $P_{h,i,j}$ and $\tau_{h,i,j}$ can be computed from a user's mobility history to help improve the accuracy of the next-cell predictions.

Traffic patterns typically exhibit some form of seasonality whereby user mobility behaviors are likely related to a certain epoch in time and repeats periodically. A good example would be the mobility pattern of students within a university campus

which tends to be consistent within a single term and varies from one academic term to another. Another example is a commuter's daily travel patterns during the week. A more accurate prediction would need to address these season-based changes in mobility patterns and compute the necessary parameters for each of those distinct periods to be applied exclusively for these periods. Notice that some instances could be the result of temporary occurrences such as road repairs which can have a significant impact on the mobility behavior of the users. Such occurrences should not be treated as being due to season changes since they are assumed to be infrequent situations. Other infrequent occurrences include severe traffic jams, weather, and geographic disasters that can be assumed to be temporary and rare. The model covered in this thesis does not consider the reaction to such uncommon occurrences.

2.2 Markov Renewal Processes

A ***M**arkov **R**enewal **P**rocess* (MRP) is a semi-Markov process where the successive state occupancies are governed by the transition probabilities $P_{i,j}$ of a Markov process, and the sojourn time in any state depends on both the current state and the next state transition. The behavior of this process is defined by the pair of random variables $\{X, T\}$. A more detailed description of such processes can be found in [70]. The semi-Markov kernel for a time-homogeneous process is given by $Q_{i,j}(t)$,

$$Q_{i,j}(t) \;=\; Pr\{X_{n+1} = j \;\;,\;\; T_{n+1} - T_n \leq t \mid X_n = i\}, \tag{2.1}$$

where X_n and X_{n+1} represent the state of the system after the n-th and $(n+1)$-th transition, respectively, with T_n and T_{n+1} being the times at which the n-th and $(n+1)$-th transitions occur, respectively. $Q_{i,j}(t)$ denotes the probability that upon entering state i, the process makes a transition into state j within t units of time of being in state i. The time t in such a process can either be discrete or continuous. We can further re-write the kernel $Q_{i,j}(t)$ as follows,

$$Q_{i,j}(t) \;=\; P_{i,j}\; G_{i,j}(t), \tag{2.2}$$

where

$$G_{i,j}(t) \;=\; Pr\{T_{n+1} - T_n \leq t \;\mid\; X_{n+1} = j \quad , \quad X_n = i\}. \tag{2.3}$$

$G_{i,j}(t)$ represents the conditional probability that a transition will take place within an amount of time t, given that the process has just entered state i and will next transition to state j. The sojourn times in such a process are assumed to follow any arbitrary distribution and allows for a convenient departure from the common assumption of exponentially distributed sojourn times [7, 71], thus permitting a more accurate representation of the temporal behaviors. Many of the mobility models developed by various researchers had tended to assume an exponentially distributed sojourn time due to it's favored "memoryless" property [72], which yielded a more tractable model. The geometric distribution possesses the same "memoryless" property in discrete-time models. However, the exponential distribution may not be representative of all types of sojourn time behaviors. An MRP with exponentially distributed sojourn times reduces to the well-known continuous-time Markov process.

Since it is known that the limits of the cumulative distribution of the sojourn times are $G_{i,j}(t) \to 1$ as $t \to \infty$, the following limits are true for the semi-Markov kernels,

$$P_{i,j} \;=\; \lim_{t\to\infty} Q_{i,j}(t). \tag{2.4}$$

Define the random variable ω_i as the time spent in the current state i before making a transition, then the cumulative waiting time probability is given as follows.

$$W_i(t) \;=\; Pr\{\omega_i \le t\} \;=\; \sum_j P_{i,j} G_{i,j}(t) \qquad \forall i \tag{2.5}$$

Other performance metrics from an MRP can be found in [76].

We can also define the kernel $q_{i,j}(t)$, such that

$$\begin{aligned} q_{i,j}(t) &= Pr\{X_{n+1} = j,\ T_{n+1} - T_n = t \mid X_n = i\} \\ &= P_{i,j}\ g_{i,j}(t), \end{aligned} \tag{2.6}$$

where $g_{i,j}(t)$ is the probability that the process will sojourn in state i for exactly t units of time before making a transition into state j, with

$$g_{i,j}(t) = Pr\{T_{n+1} - T_n = t \mid X_{n+1} = j,\ X_n = i\}. \tag{2.7}$$

Phase-type distributions are a class of probability distributions that are used to approximate any positive valued distributions. Such distributions can be represented by a single random variable which describes the time until absorption of a finite Markov process with a transition probability/rate matrix $\mathbf{S}$ and one absorbing state, with $\boldsymbol{\alpha}$ being the probability vector governing the starting state of the process. Neuts in [73] provides a detailed description of phase-type distributions that are typically characterized as $(\boldsymbol{\alpha}, \mathbf{S})$. A phase-type distribution of $(\boldsymbol{\alpha}, \mathbf{S})$ could be used to model the state sojourn times $G_{i,j}(t)$. If we assume that the sojourn times in state i for each $i \rightarrow j$ transition of the MRP have a phase-type distribution $(\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ in

continuous-time, then the elements $G_{i,j}(t)$ will have the following form,

$$G_{i,j}(t) \;=\; 1 - \boldsymbol{\alpha}_{i,j} \exp\left(\mathbf{S}_{i,j} t\right) \mathbf{1} \qquad \text{for} \;\; t \geq 0, \tag{2.8}$$

where $\mathbf{1}$ is a column vector of ones. The mean of such distributions is given as $-\boldsymbol{\alpha}_{i,j}\mathbf{S}_{i,j}^{-1}\mathbf{1}$. The probability density function of such distributions in continuous-time is given by $g_{i,j}(t) = \boldsymbol{\alpha}_{i,j} \exp\left(\mathbf{S}_{i,j} t\right) \mathbf{S}_{i,j}^{0}$, where $\mathbf{S}_{i,j}^{0} = -\mathbf{S}_{i,j}\mathbf{1}$.

If $G_{i,j}(t)$ assumes a discrete-time phase-type distribution, then it will be of the following form,

$$G_{i,j}(t) \;=\; 1 \;-\; \boldsymbol{\alpha}_{i,j}\mathbf{S}_{i,j}^{t}\mathbf{1} \qquad \text{for} \;\; t = 0, 1, 2, \cdots, \tag{2.9}$$

which has a mean of $\boldsymbol{\alpha}_{i,j}\left(I - \mathbf{S}_{i,j}\right)^{-1}\mathbf{1}$, where I is an identity matrix with the same dimensions as $\mathbf{S}_{i,j}$. The probability density function of such distributions in discrete-time is given by $g_{i,j}(t) = \boldsymbol{\alpha}_{i,j}\mathbf{S}_{i,j}^{t-1}\mathbf{S}_{i,j}^{0}$, where $\mathbf{S}_{i,j}^{0} = \mathbf{1} - \mathbf{S}_{i,j}\mathbf{1}$. The fitting of data to phase-type distributions have been covered by various researchers, such as Horváth and Telek in [74], and Panchenko and Thümmler in [75].

2.3 Model Description

The mobility behavior can be modeled as a Markov renewal process and can be used to predict the transition that an arbitrary user makes from its current location, within a time t (and not *at* time t, for a better prediction as argued by Song et al. in [28]). The model assumes the availability of the information regarding the location transition probabilities and the conditional distributions of the cell-sojourn times, using the aggregate mobility history that is collected in each cell of the network. The probability $Q_{i,j}(t)$ defined in Equation (2.1) can be computed to evaluate the predictions of an arbitrary user making a transition to a neighboring location, which depends on the length of time spent in the current location.

The majority of the mobility models proposed in the literature focused only on those users in a network with an *active* session. However, a more elaborate model could consider distinguishing the users that are mobile with active and idle sessions [77], as well as the changes in the session conditions (i.e. from active to idle and vice versa). A mobile user is said to have an idle session when his/her wireless terminal is not engaged in any transmission/reception and is on standby. In [78], Sricharan and Vaidehi argued that the lack of consideration of the idle mode behavior of the mobile users can have a significant impact on several network management tasks. Halepovic and Williamson in [79] had examined some traffic traces and discovered that there exists some correlation between the call activity and mobility patterns for the users.

The proposed mobility prediction scheme is based on the assumption that a user's session-activity patterns are correlated with the location and movement patterns in a network. In addition, the presence of users with idle sessions may exert some influence on the density and behavior of users with active sessions, and vice versa. In fact, mobile users with idle sessions can potentially initiate a network access, especially when receiving a transmission from the network, e.g. a mobile user receiving a request to answer an incoming call in cellular networks. Various networking technologies do also keep track of their registered users that have no ongoing sessions (i.e. idle sessions) for "paging" purposes [80]. For example, mobile users with idle sessions in cellular networks generate location updates while roaming, since the network needs to be able to determine where to direct an incoming call for a particular user. In WLANs, a user with an idle session remains associated with the network and re-associates its connection with the network when moving from the coverage area of one access point to another.

For our prediction scheme, we propose to define the transition probabilities as follows.

- $P_{i,j}$ denotes the probability that a user's ongoing and active session is transferred from cell i to cell j.

- $P_{i,-i}$ denotes the probability that a user's ongoing and active session is terminated in the current cell i.

- $P_{-i,i}$ denotes the probability that a user's idle connection becomes active in

the current cell i.

- $P_{-i,-j}$ denotes the probability that a user's idle connection is transferred from cell i to cell j.

The subscripts i and $-i$ denote a user in location i with an active and idle session, respectively. It is further assumed that $P_{i,i} = 0$ and $P_{-i,-i} = 0$ since this model describes a "renewal" process for predicting the future transitions. Furthermore, in many cases it can be assumed that a change in location and session-activity cannot occur simultaneously in a single transition if we were only to consider the cases when a "renewal" in either of the states occur. This is true for the case where the mobility model is constructed in continuous-time. However, the specification of the model can be readily adjusted to account for the simultaneous transitions in both location and session-activity. The former assumption was chosen to be adopted throughout this thesis.

For each of the probabilities given above, we can define a cumulative distribution function in the form given by Equation (2.3), and subsequently compute the semi-Markov kernel $Q_{i,j}(t)$. In general, the empirical distribution can be directly used, or these distributions can assume any closed-form distribution function. These closed-form functions can be obtained by passing the sojourn times collected from the traffic traces through some distribution fitting tools such as [81] and [82] that attempt to best fit the data to some known distribution function. Another approach would be to try and fit the data to a phase-type distribution using the various methods given

in the literature such as the one proposed in [74] and [75]. Other distribution fitting methods can also be used provided that a distribution function can be formulated in the form given by Equation (2.3). Note that the proposed mobility model is not limited to any particular types of sojourn time distributions, but the accuracy of the model (and ultimately the predictions) will be greatly influenced by how well the sojourn times data can be fitted to an appropriate distribution.

Let X be a random variable that defines the state of a mobile user in terms of its location and session activity, such that $X \in \mathcal{X}$. Let L_I and L_A be the total number of cells/locations with users having idle and active network sessions, respectively. Thus, X has the following state space $\mathcal{X} = \{(-1, -2, \cdots, -L_I)\ ,\ (1, 2, \cdots, L_A)\}$, where typically $L_A = L_I$, but generally $L_A \geq L_I$. Note that the states $X = -i$ and $X = i$ defines the presence of a user in location i with idle and active sessions, respectively. For illustrative purposes, let us assume a simple example of the 7-cell structure in Figure 2.1 as being the entire network coverage area. An MRP model that considers the explicit cell transitions while a user is engaged in both an active or idle network session can be formulated with the state space given by $\mathcal{X} = \{(-1, -2, -3, -4, -5, -6, -7)\ ,\ (1, 2, 3, 4, 5, 6, 7)\}$, where $L_I = L_A = 7$. In this case, the probability of having a session terminated in cell 7 within time t is given by the element $Q_{7,-7}(t)$ and the remaining set of possible transitions can be explained in the same manner.

Not all networks track their users' exact cell changes while their sessions are

idle. For example, cellular networks only monitor changes in a user's "location area", which is equivalent to a certain cluster of cells, while the session is idle. Our model can be applied to such circumstances by appropriately re-defining the subset of states assigned for those users at the various locations that have idle sessions. Hence, L_A can also be defined as the total number of cells and L_I is the total number of "location areas" with $L_A \neq L_I$. For cellular networks, $L_A \geq L_I$ whereas it is common to find that $L_A = L_I$ in WLAN. Using the same 7-cell example given in Figure 2.1, the network may impose a limited tracking of their users with idle sessions based on their "location areas" where, as an example, these areas are defined such that cells $(1, 4, 7)$ are labeled as $X = -1$, cells $(2, 3)$ are labeled as $X = -2$, and cells $(5, 6)$ are labeled as $X = -3$. Thus, the state space for this MRP model reduces to $\mathcal{X} = \{(-1, -2, -3)\ ,\ (1, 2, 3, 4, 5, 6, 7)\}$, whereby a transition that reflects the termination of a session in cell 7 is now given by the element $Q_{7,-1}(t)$. Note that in such cases, a transition from cell 7 to cell 1 with an idle session is not considered in this reduced MRP model since the user in actuality remains in the same location area without any change in state. However, a transition of the same idle session from cell 7 to cell 6 does involve a change in state from -1 to -3.

The other extreme circumstance is if the network does not monitor nor consider the state of a mobile user with an idle session in its predictions. In such cases, the network would commence its predictions using the MRP model with state space $\mathcal{X} = \{1, 2, 3, 4, 5, 6, 7\}$ once a user's session is initiated and continue to assume that

the connection will remain active until terminated by the user. Hence, our proposed MRP-based prediction scheme can be tailored to the needs and requirements of different networks.

The kernels $Q_{i,j}(t)$ could further be used to construct a semi-Markov kernel matrix $\mathbf{Q}(t) = \{Q_{i,j}(t)\}$. Assume the simple example of a 7-cell network given in Figure 2.1 whereby users in cell 1 can transition directly into the six other surrounding cells, while users in cells 2 to 7 can only make a direct transition into three other neighboring cells. For this 7-cell example, the matrix $\mathbf{Q}(t)$ can be constructed as follows, with the order of the state space give as $\{-1, -2, \cdots, -7, 1, 2, \cdots, 7\}$.

$$\mathbf{Q(t)} \;=\; \begin{pmatrix} \mathbf{M}^{-}(t) & \mathbf{S}^{+}(t) \\ \mathbf{S}^{-}(t) & \mathbf{M}^{+}(t) \end{pmatrix}, \qquad \text{where} \tag{2.10}$$

$$\mathbf{M}^{+}(t) \;=\; \begin{pmatrix} 0 & Q_{1,2}(t) & Q_{1,3}(t) & Q_{1,4}(t) & Q_{1,5}(t) & Q_{1,6}(t) & Q_{1,7}(t) \\ Q_{2,1}(t) & 0 & Q_{2,3}(t) & 0 & 0 & 0 & Q_{2,7}(t)) \\ Q_{3,1}(t) & Q_{3,2}(t) & 0 & Q_{3,4}(t) & 0 & 0 & 0 \\ Q_{4,1}(t) & 0 & Q_{4,3}(t) & 0 & Q_{4,5}(t) & 0 & 0 \\ Q_{5,1}(t) & 0 & 0 & Q_{5,4}(t) & 0 & Q_{5,6}(t) & 0 \\ Q_{6,1}(t) & 0 & 0 & 0 & Q_{6,5}(t) & 0 & Q_{6,7}(t) \\ Q_{7,1}(t) & Q_{7,2}(t) & 0 & 0 & 0 & Q_{7,6}(t) & 0 \end{pmatrix}, \tag{2.11}$$

$$\mathbf{M}^{-}(t) \;=\; \begin{pmatrix} 0 & Q_{-1,-2}(t) & Q_{-1,-3}(t) & Q_{-1,-4}(t) & Q_{-1,-5}(t) & Q_{-1,-6}(t) & Q_{-1,-7}(t) \\ Q_{-2,-1}(t) & 0 & Q_{-2,-3}(t) & 0 & 0 & 0 & Q_{-2,-7}(t)) \\ Q_{-3,-1}(t) & Q_{-3,-2}(t) & 0 & Q_{-3,-4}(t) & 0 & 0 & 0 \\ Q_{-4,-1}(t) & 0 & Q_{-4,-3}(t) & 0 & Q_{-4,-5}(t) & 0 & 0 \\ Q_{-5,-1}(t) & 0 & 0 & Q_{-5,-4}(t) & 0 & Q_{-5,-6}(t) & 0 \\ Q_{-6,-1}(t) & 0 & 0 & 0 & Q_{-6,-5}(t) & 0 & Q_{-6,-7}(t) \\ Q_{-7,-1}(t) & Q_{-7,-2}(t) & 0 & 0 & 0 & Q_{-7,-6}(t) & 0 \end{pmatrix}, \tag{2.12}$$

$$\mathbf{S}^{+}(t) \;=\; diag\left[Q_{-1,1}(t), Q_{-2,2}(t), Q_{-3,3}(t), Q_{-4,4}(t), Q_{-5,5}(t), Q_{-6,6}(t), Q_{-7,7}(t)\right], \tag{2.13}$$

$$\mathbf{S}^{-}(t) \;=\; diag\left[Q_{1,-1}(t), Q_{2,-2}(t), Q_{3,-3}(t), Q_{4,-4}(t), Q_{5,-5}(t), Q_{6,-6}(t), Q_{7,-7}(t)\right], \tag{2.14}$$

with $diag\left[\cdots\right]$ being a square matrix of zeros with the elements $[\cdots]$ along its diagonal. The elements in $\mathbf{M}^{+}(t)$ represent the transitions that correspond to a change in location from one cell to another due to mobility and while running an active session, while the elements in $\mathbf{M}^{-}(t)$ are for the case where the location transitions involve users with idle network sessions. These transitions only include those made to the cells that directly neighbor those from which the transitions are made, e.g. $Q_{6,3}(t) = 0$ since cell 6 is not a direct neighbor of cell 3. The elements in $\mathbf{S}^{+}(t)$ represent the transitions involving a user's session activity changing from idle to active (i.e., network session initiation) while residing in the same cell, whereas $\mathbf{S}^{-}(t)$ are for the case of session activities changing from active to idle (i.e., network session completion). Note how all these transitions do not include those that involve changes in both location and session activity simultaneously. This example models for mobile users with both active and idle network sessions. If the focus need only be on those users with active sessions alone, then the semi-Markov kernel matrix simply reduces to the one given by $\mathbf{M}^{+}(t)$, i.e. $\mathbf{Q}(t) = \mathbf{M}^{+}(t)$ for the case of active users alone in the 7-cell network example.

The Markov renewal process for mobility prediction can also be extended to the case where the user's previous location is considered in the mobility pattern, i.e. extending the kernels $Q_{i,j}(t)$ to $Q_{h,i,j}(t)$ which has the following form.

$$\begin{aligned} Q_{h,i,j}(t) &= Pr\{X_{n+1} = j \qquad T_{n+1} - T_n \leq t \mid X_n = i \;\;,\;\; X_{n-1} = h\} \\ &= P_{h,i,j}\; G_{h,i,j}(t). \end{aligned} \tag{2.15}$$

It has been suggested by Choi and Shin [22], as well as several other researchers, that estimating the likelihood of future transitions *within a certain time window* can yield better predictions as opposed to predicting the time at which a transition could occur. Given our definition for $Q_{i,j}(t)$, we can further compute

$$Q_{i,j}(t_b, t_f) \;=\; Q_{i,j}(t_f) \;-\; Q_{i,j}(t_b), \tag{2.16}$$

where $Q_{i,j}(t_b, t_f)$ is the probability of observing a transition into state j from state i within a time period t, such that $t_b < t \leq t_f$. The size of the time window (t_b, t_f) can have an influence on the prediction accuracy and thus could be dynamically "fine-tuned" according to some criteria.

2.4 Prediction and Confidence

The results returned by the elements $Q_{i,j}(t)$ can generally be applied for numerous purposes including the prediction of an arbitrary user's mobility behavior. They can further be used for other benefits such as estimating the future resource allocations and understanding the spatial-temporal traffic demands. Various uses will be presented in this thesis for the purpose of demonstrating the benefits of the proposed model. This section describes the approach used to evaluate the predictions from the mobility model specified in the previous section.

Given the probabilities $Q_{i,k}(t)$, one form of prediction is to estimate the proportion of users from the current population V_i that are expected to transition into state k within time t after a single transition from state i. Hence, from our prediction results, we would expect that $V_i Q_{i,k}(t)$ from the current population V_i would end up in state k within t units. This can be construed as having made an aggregate prediction on all the V_i users that are currently in state i and how they are expected to spread amongst the various locations within time t. However, in this approach the predictions are not specific to any user in particular since the result only decides on the number of users that are expected to transition from one state to another.

In the case of a per-user prediction, the future outcome could be randomly predicted and based on the distribution defined by the transition probabilities. However, it is common to make an outcome prediction based on the maximum probability from the distribution of the set of possible events (see [20]). For example, if $P_{i,j} = \max_{k \in \Omega_i} \{P_{i,k}\}$, where $P_{i,k}$ is the transition probability from state i to k and Ω_i are the set of possible states that can be transitioned into from state i, then state j is said to be the most likely outcome to follow state i. Therefore, state j is said to be the predicted outcome for users that are expected to make a transition from

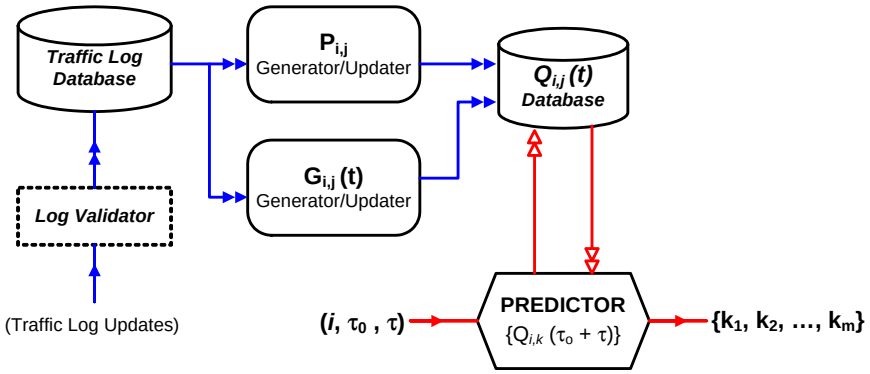


Fig. 2.4: The MRP-based prediction procedure.

state i. Some researchers had further explored the benefits of selecting the top m most likely events as the result of their predictions, such as Chan and Seneviratne in [21], with $m = 1$ being the most common choice.

The methods outlined earlier can also be applied for computing the next state predictions from the proposed semi-Markov mobility model. Hence, the time-dependent prediction results will be such that

$$\begin{aligned} Q_{i,j}(t) &= \max_{k\in\Omega_i} \{Q_{i,k}(t)\}\,, \\ \text{or} \quad q_{i,j}(t) &= \max_{k\in\Omega_i} \{q_{i,k}(t)\}\,. \end{aligned} \tag{2.17}$$

The block diagram in Figure 2.4 illustrates the basic sequence of procedures that are needed to implement the proposed MRP-based predictor. The traffic log database holds a record of the aggregate traffic logs of the users in the network. These logs contain the relevant information needed to compute the state transition probabilities $P_{i,j}$ and the state sojourn time distributions $G_{i,j}(t)$, for all valid (i, j) pairs of transitions. These components can be updated periodically, based on new logs that are received by the network and updated in the database as a result of some completed event by a mobile user. The "Log Validator" component could be

added to filter out those logs that do not accurately reflect the mobility behavior of the user, especially at the instances where a session was denied a natural completion by the network, e.g. the dropping of a session due to insufficient resources at the newly visited access point. The $Q_{i,j}(t)$ database keeps a record of the computed elements which are directly queried by the predictor for evaluating the probabilities. The predictor takes as its inputs the current state/location i, the current sojourn time τ_0 in state i, and the time length τ from the current time τ_0 during which a transition is expected to be made. Given those inputs, the predictor queries the $Q_{i,j}(t)$ database to compute the values $Q_{i,k}(\tau_0 + \tau)$ for all $k \in \Omega_i$ and outputs k_1 as the next most likely outcome to occur τ time units from the current time (which conforms with Equation (2.17)). The results of the second and up to the m-th most likely outcomes at the next transition are also returned, if needed.

In some instances, selecting the next likely outcome as the one with the maximum probability may not necessarily yield a confident prediction. For example, consider the simple case where the probabilities of going from state 1 to 2 and 1 to 3 are $P_{1,2} = 0.51$ and $P_{1,3} = 0.49$, respectively. We would expect most prediction algorithms to select state 2 as being the next likely outcome, but the probability $P_{1,3}$ is very close to $P_{1,2}$. However, if these transition probabilities were instead $P_{1,2} = 0.8$ and $P_{1,3} = 0.2$, then we would be a lot more confident in making the prediction of state 2 being the next state as opposed to the previous case. Hence, it is assumed that the higher the maximum probability is in comparison with the other transition probabilities, the more confident one can be with making a prediction using this set of probabilities.

A simple measure of confidence would be to find the entropy of the prediction

made from state i, which can be calculated as follows,

$$H\left(Q_i(t)\right) \;=\; -\sum_{k\in\Omega_i} \frac{Q_{i,k}(t)}{S_i(t)} \log\left(\frac{Q_{i,k}(t)}{S_i(t)}\right), \tag{2.18}$$

$$\text{where} \quad S_i(t) \;=\; \sum_{k\in\Omega_i} Q_{i,k}(t),$$

$$\text{such that} \quad 0 \;\leq\; H\left(Q_i(t)\right) \;\leq\; \log\left(|\Omega_i|\right),$$

with $|\Omega_i|$ being the number of possible outcomes from state i and $S_i(t)$ is needed to normalize our semi-Markov transitional probabilities, since $S_i(t) \leq 1$ and $S_i(t) \to 1$ as $t \to \infty$. A similar definition can be obtained for the case of using $q_{i,j}(t)$. Note that the measure of prediction confidence is time-dependent and the higher the $H\left(Q_i(t)\right)$, the less confident one may be of the prediction. Therefore, one way of increasing the confidence of the prediction is to vary the time t within which the transition is expected to be made. Increasing t will not necessarily decrease $H\left(Q_i(t)\right)$ since it is strongly dependent on the shape of the state sojourn time distribution. Another approach for decreasing $H\left(Q_i(t)\right)$ would be to lump together the top m outcomes with the highest probabilities into a single probability and recompute the new confidence measure, for predicting a cluster of m states as the next outcome.

3. PRACTICAL EVALUATIONS: WLAN SCENARIO

To illustrate the behavior and benefits of the proposed mobility prediction, a set of actual wireless traffic traces were used for computing the parameters needed to make a prediction, i.e. the transition probabilities and the cell sojourn times. The data set was acquired from the CRAWDAD (Community Resource for Archiving Wireless Data at Dartmouth) online repository [83] which included the recording of WLAN traffic logs produced by 623 wireless access points at Dartmouth College (New Hampshire-USA) from the period of April 2001 until June 2004. The set contained 13888 logs, one for each unique wireless terminal that acquired access to the network during the monitored period. Within the logs, details of the access point ID as well as the time at which the user associated its terminal with the access point were included. This allowed for computing the location transition probabilities as well as the times spent at each location before a transition to the next location. A snapshot of one of those logs is given in Figure 3.1 for the user with MAC Address 000423de1f86. Each entry in the logs contained the time (first column) at which

1076098966	AcadBldg1AP3
1076099208	ResBldg97AP5
1076099464	ResBldg97AP2
1076100005	OFF
1076100005	ResBldg97AP5
1076100697	ResBldg97AP2
1076100940	OFF
1076100940	ResBldg97AP5
1076101322	AcadBldg1AP3

Fig. 3.1: Part of a movement log from the CRAWDAD data.

the user first associated its connection with the access point and the second column gives the unique ID of that particular access point. The "OFF" refers to the case where the user terminated its session at the same access point that was previously logged. The event time in the first column corresponds to the number of seconds elapsed since January 1st 1970 00 : 00 GMT. These logs for each mobile terminal with a unique MAC address were actually extracted from a much larger trace set that contained continuous recordings of the syslog records generated by the access points. There were a few events that showed some discontinuity in the records due to some minor technical issues that had been identified by the researchers who collected the traffic traces. Such occurrences were not included in the analysis. Each syslog event contained the information needed to generate the movement logs for each user such as *timestamp*, *access_point_ID*, *MAC_address*, and *syslog_message* indicating the changes in the association between the network and the user. Figure 3.2 is an example of some syslog entries that were taken from the trace set.

```
986996241 Apr 11 09:37:21   AcadBldg33AP6   (Info): Station 004096daa8fe Authenticated
986996241 Apr 11 09:37:21   AcadBldg33AP6   (Info): Station 004096daa8fe Associated
986996363 Apr 11 09:39:23   AcadBldg33AP5   (Info): Station 00409630cdc9 roamed
986996363 Apr 11 09:39:23   AcadBldg33AP5   (Info): Station 00409630cdc9 roamed
986996363 Apr 11 09:39:23   AcadBldg33AP6   (Info): Station 00409630cdc9 Authenticated
986996363 Apr 11 09:39:23   AcadBldg33AP6   (Info): Station 00409630cdc9 Reassociated
986996680 Apr 11 09:44:40   AdmBldg19AP3    (Info): Station 0040961e58be Reassociated
```

Fig. 3.2: Part of a syslog trace collected on April 11th 2001.

The traces in the data set do not show the details of the traffic flow between the user's wireless terminal and the network and instead focuses on the details of the time and place at which the user's terminal associates, transfers, and terminates its session with the network. In other words, the user's terminal may not always be engaged in any traffic flow with the network during the entire period of time that it is associated with a particular access point.

When computing the parameters for $Q_{i,j}(t)$, an approach that is similar to the

one suggested by Lee and Hou in [66] was used for filtering out any events that display frequent re-associations between a group of access points within a short period of time. Such behaviors are known as *ping-pong* transitions. These type of transitions have been shown to have a significant impact on the computation of the parameters needed for the mobility predictions. Most events that had lasted less than 30 seconds had also been involved in instances with frequent re-associations and were further filtered out from the analysis.

Both Matlab© and Flanagan's Java scientific library [81] were further employed for extracting the best possible fit for the sojourn time distributions. Any event that had lasted less than 30 seconds or more than 5 hours were ignored in the analysis. A considerable number of events in the data set had lasted for very long periods of time, some of which showed a continuous connection with the WLAN for more than a day. There was no information to suggest whether such terminals are static, e.g. desktop computers, in which case they would unlikely be involved in any mobile activities. However, such behaviors may be expected in WLAN type infrastructures and the sojourn times could be appropriately fitted to some heavy-tailed distribution. The data set did not have enough logs with such events to generate a good fit and thus it was chosen to omit these events from the model parameter computations, as well as the prediction results. These omissions were chosen to be done in the attempt to limit the influence of the errors encountered in the distribution fitting process on the accuracy of the predictions returned by the proposed model.

3.1 Analysis of Mobility Behaviors

Before proceeding with analyzing the effectiveness of the proposed prediction scheme, the ability to take a standard data set of traffic logs and use it to compute the required parameters of the model will first be demonstrated. We then illustrate how these results can be used to model the mobility behavior of an arbitrary user in the network as well as give examples on how to interpret some of those results. We are aware that there are various mobility models that are better at capturing the true roaming behavior of its users. However, most of them rely on the availability of extra information that may or may not be readily available in the standard traffic logs. The aim here is to only use the data that is available in many of the standard logs and investigate how much information can be inferred from it without resorting to other types of information found elsewhere. The examples given in this section will help show how the proposed model could also be used to understand the behavior of the mobile users, along with the prediction of future mobility transitions.

In the numerical evaluations, we focused only on a subset of the entire data set, namely those events that exhibit transitions between the 21 different access points, each served by their own wireless access point, within the Library Building 2 of the Dartmouth College campus. The traffic traces provided by [83] also supplied the details of where the wireless access points are deployed across the Dartmouth College campus. The information included both the coordinates and the floor number of where these access points are located in the Library Building 2 which was collected from the AutoCAD drawings of the premises. An outline of where these 21 access points are located in the building is shown in Figure 3.3, as viewed from the front side of the Library Building 2. The figure also shows how these access points are distributed across the various floors of the building. It was assumed that their

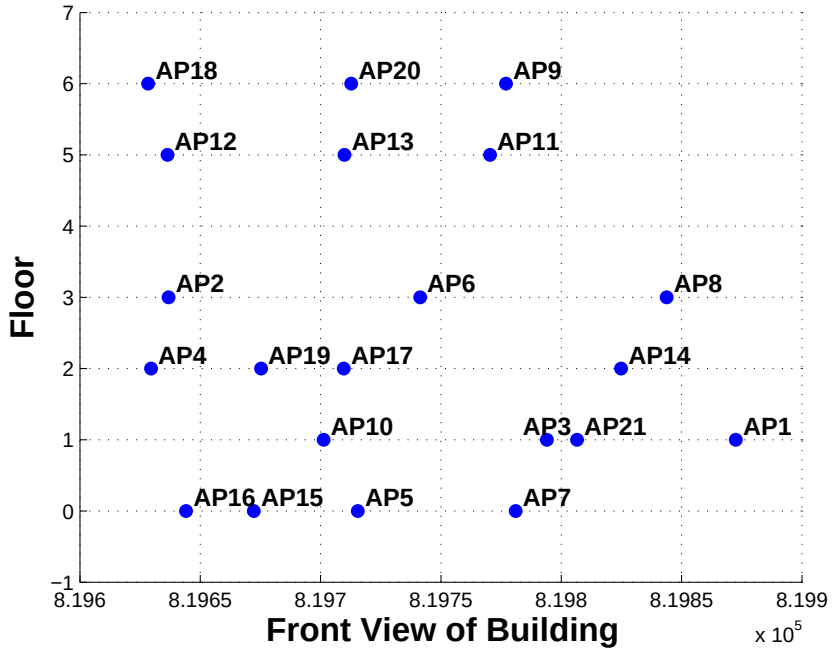


Fig. 3.3: A front view of the spatial locations of the 21 wireless access points in Library Building 2 at Dartmouth College.

positions remained to be the same throughout the data traffic collection process. The selection of the Library Building 2 was driven by the fact there were a large number of events across the period of 3 years in that particular subset which provided enough data for computing the required parameters.

When examining the sojourn times, the two-parameter Log-normal distribution was found to give the best fit amongst the list of various distributions that are include in Matlab's Distribution Fitting Toolbox. The best fit was determined by the Log likelihood metric that was computed in the toolbox during the fitting process. The two-parameter Log-normal distribution was chosen to be used for evaluating the elements in the semi-Markov model.

Figure 3.4 and Figure 3.5 are plots of the probabilities $Q_{14,j}(t)$ and $Q_{17,j}(t)$, respectively, where the next state j is from the set of possible locations that a user

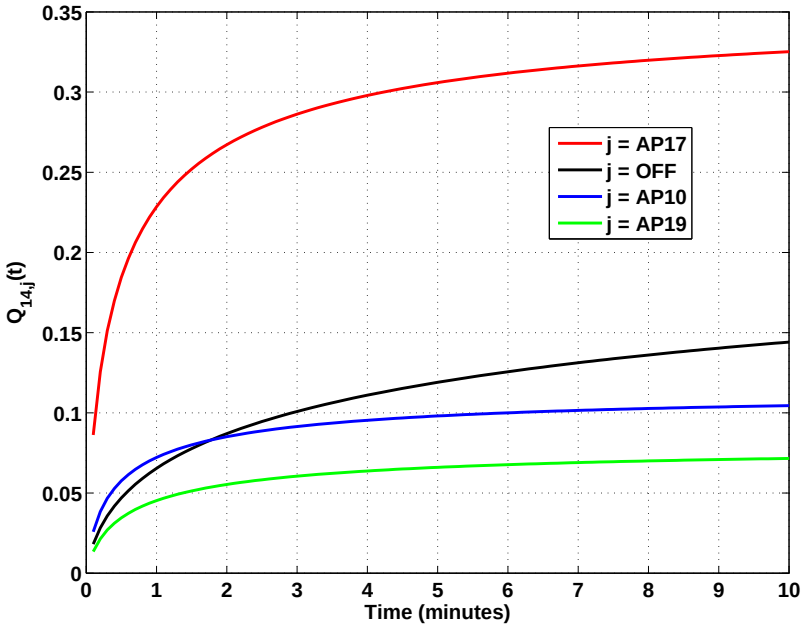


Fig. 3.4: A plot of the time-varying probabilities of an arbitrary user making a transition from AP14.

can transition into and from the current location, with $j = OFF$ being the state where the user disassociates its connection with the network at the current location. Both graphs only show the top 4 next state transitions with the highest probabilities and how they vary at different periods of time. For example, Figure 3.4 shows that a user is most likely to make his/her next transition into AP17 from AP14. It is also shows that an arbitrary user is next more likely to make a transition into AP10 when compared with terminating his/her network session, within the first 2 minutes of initially being connected to AP14. However, if the same user remains connected to AP14 for more than 2 minutes, then he/she will instead be next more likely to terminate their network session in the current location rather than transitioning to AP10.

One way a network provider might make use of such information is in deciding *when*, *where*, and for *how long* a certain amount of resource would need to be

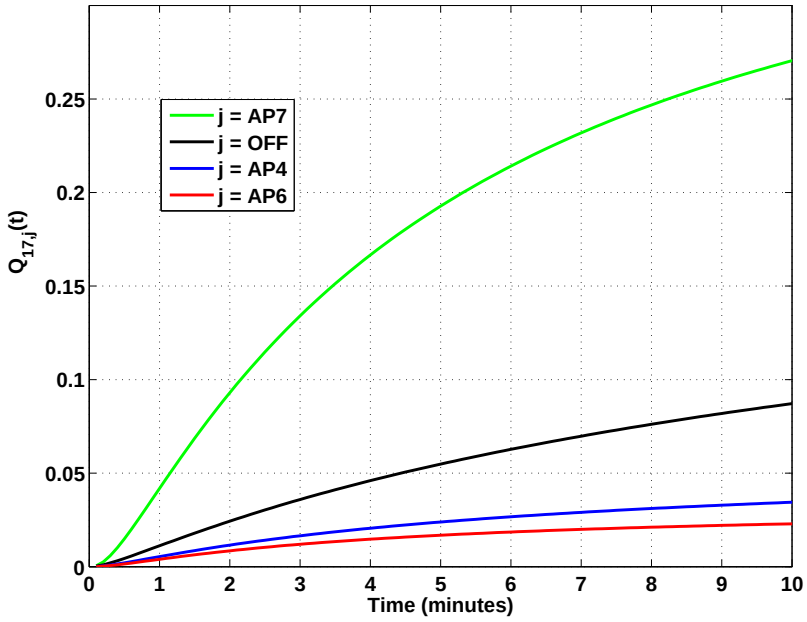


Fig. 3.5: A plot of the time-varying probabilities of an arbitrary user making a transition from AP17.

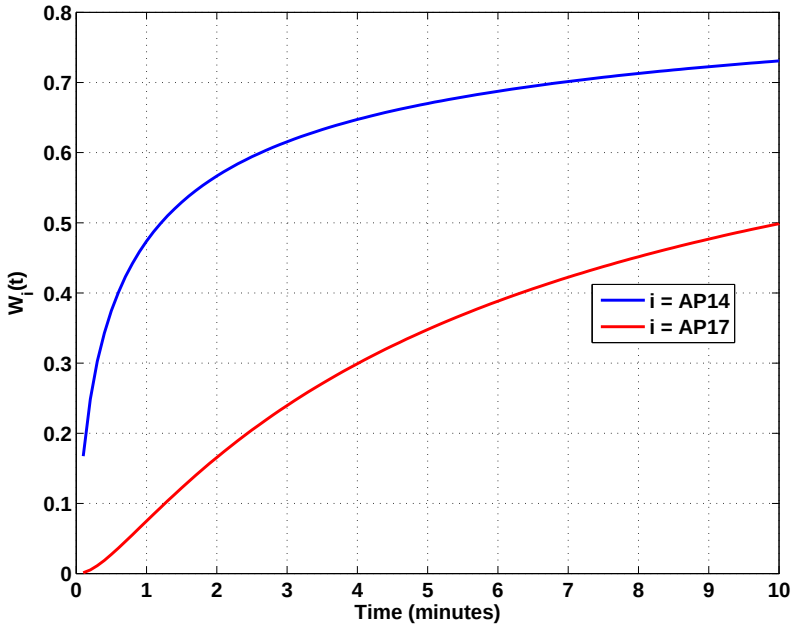


Fig. 3.6: A plot of the sojourn times in state i.

reserved for each user in order to ensure the successful transfer of an active session between the access points. If we consider again the case of a user that has a session associated with AP14, the necessary amount of resources could be initially reserved at AP17 and AP10 during the first 2 minutes of the session time or until the session is terminated in the current location, whichever is less. This assumes that the network considers only the first two most likely transitions in their prediction, as an example. If the session remains active after 2 minutes without any transitions, then this could prompt the network to limit its resource reservation thereafter to AP17 since the second most likely transition is the user terminating its session in the current location.

The network manager could come across a situation that includes multiple transitions having equal probabilities, as identified by Song et al. in [32]. An example of such an occurrence is where the two plots for $j = AP10$ and $j = OFF$ cross with each other in Figure 3.4. Such instances are more likely to be encountered when using transition probabilities alone to compute the predictions. To deal with such incidents, a tie-breaking method would need to be implemented. A simple approach would be to either consider *all* transitions with equal probabilities as the result of the prediction computation, or select a certain subset of these transitions at random. An alternative method would be to monitor the rising rate of the $Q_{i,j}(t)$ function and select the one with the highest rate as the prediction result, since the higher rate is assumed to be leading towards a higher probability. For example, at the time instant of just under 2 minutes in Figure 3.4 where $Q_{14,10}(t) = Q_{14,OFF}(t)$, the transition $AP14 \rightarrow OFF$ will be the chosen prediction under the third proposed tie-breaker policy, due to a higher rate of increase for $Q_{14,10}(t)$.

Figure 3.6 shows the probabilities of an arbitrary user's session sojourn time in the current location before making any transition out of the current location for

AP14 and AP17, including the session termination in the current location. These results show that an arbitrary user is more likely to associate its session with AP14 for less time when compared with those associated with AP17. This could imply that users within the vicinity of AP14 tend to be more mobile (from the network's perspective) or tend to complete their sessions much sooner than those being served by AP17. The network could utilize this type of information to determine *when* it should start computing the mobility predictions of its users. One possible course of action would be to assign a threshold to decide how long the network should wait before initiating any prediction computations. They may help with reducing the amount of computations and predictions generated by the network manager. For example, using the results in Figure 3.6, the network could consider assigning a probability threshold of $p = 0.3$ which would suggest that the mobility predictions for users associated with AP17 need only commence after 4 minutes.

Using Equation (2.18), the time-varying levels of prediction confidence for both cases of transitioning out of AP14 and AP17 were computed and the results are shown in Figures 3.7 and 3.8, respectively. The flat-line graph (i.e. $H_i(Q(\infty))$) included in both plots corresponds to the case of making predictions using the transition probabilities alone, which are independent of time. Moreover, the maximum entropy in both cases are $\max_{t>0}\{H(Q_{14}(t))\} = 2.31$ and $\max_{t>0}\{H(Q_{17}(t))\} = 2.29$, which occur at $t \approx 11$ minutes and $t \approx 0.1$ minute, respectively. A closer look at the results from the example given in Figure 3.8 reveals that predictions made within time $t = 0.2$ and $t = 10$ minutes may be relatively higher in confidence when compared with those obtained by using the transition probabilities alone. Hence, is it assumed that the lower the entropy, the better the level of confidence in the predictions, as defined in Section 2.4. Notice that the confidence levels of our MRP model will eventually tend to the result given by $H(Q_i(\infty))$ since $Q_{i,j}(t) \to P_{i,j}$ as $t \to \infty$.

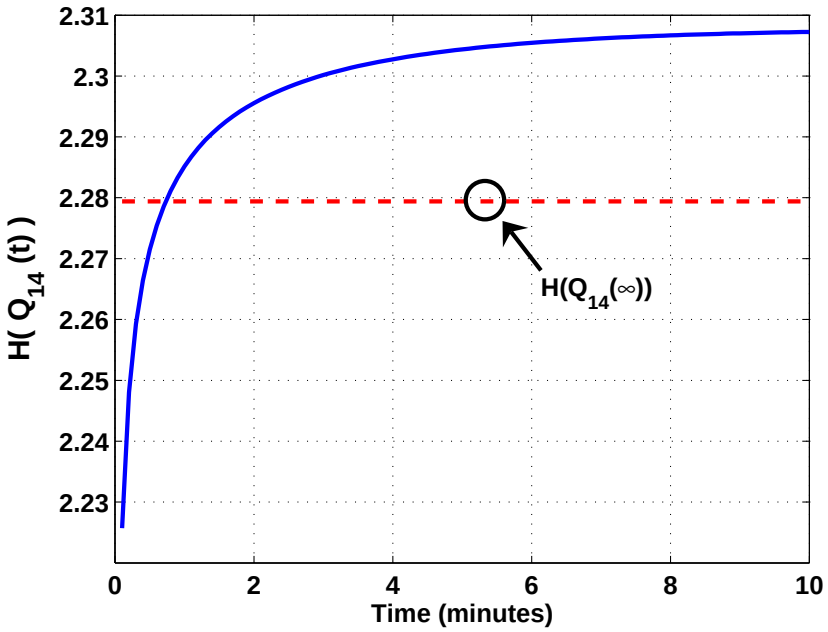


Fig. 3.7: A plot of the time-varying prediction confidence for $Q_{14,j}(t)$.

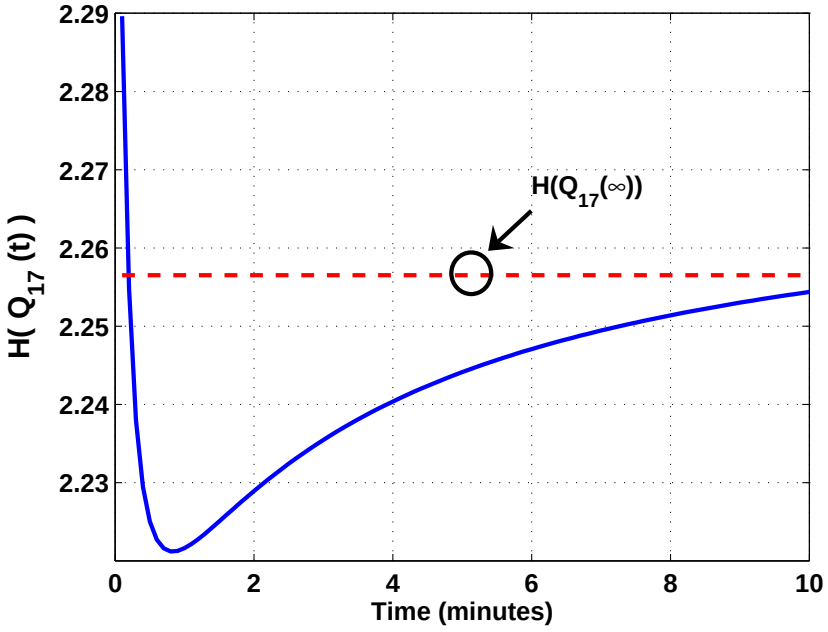


Fig. 3.8: A plot of the time-varying prediction confidence for $Q_{17,j}(t)$.

These levels of confidence correspond to how wide the difference is between the results $Q_{i,j}(t)$ for a given state i and within a time period of t. The variations in the levels shown in both Figures 3.7 and 3.8 can be understood by re-examining the results of the kernel behaviors shown in Figures 3.4 and 3.5, respectively. In Figure 3.5, there is a significantly wider margin of difference between the result for $Q_{17,7}(t)$ and the rest of the elements. The difference between the probabilities shown in Figure 3.5 becomes smaller as t increases. This might explain the increase in the entropy shown in Figure 3.8. A smaller margin of difference is observed between the result for $Q_{14,17}(t)$ and the rest of the elements in Figure 3.4. It was further observed that as t increases the levels of confidence for both results decrease (level of uncertainty increases) beyond the edge of $H\left(Q_i(\infty)\right)$ before settling towards it. For the plot shown in Figure 3.8, this reduction in confidence level occurs after $t = 12$ minutes.

Figure 3.8 shows that the confidence levels in the predictions are relatively higher when applying the proposed MRP model in comparison to using the transition probabilities alone, as reflected by the lower entropy. This is true for predictions that involve transitions to be expected within 10 minutes of being associated with the current access point. However, the result does not imply that the predictions generated by the MRP model will necessarily be more accurate. On the other hand, Figure 3.7 seems to suggest the opposite and that the confidence levels are considerably less when choosing to apply the MRP model. The proposed model has the advantage of incorporating temporal influences and could still yield a more accurate prediction than those employing the transition probabilities alone. These confidence levels may also suggest that selecting a single state/location as the prediction result at the current time period could be insufficient due to the closeness of the computed probabilities $Q_{i,j}(t)$. This information could guide the predictor to instead consider

returning the next m (with $m \geq 2$) most likely states/locations in its prediction results, where the magnitude of m could be based on some assigned threshold on the confidence levels.

3.2 Prediction Results and Accuracy

In this section, we examine the accuracy of the proposed mobility prediction and compare it with those prediction schemes that are limited to using the mobility history alone and the current state of the user. As an example, we continued to focus the attention on those transitions that are made within Library Building 2 and between 21 access points. A subset of the same traffic trace [83], namely those logs recorded from September 2001 until September 2003, were processed for computing the required MRP parameters. This subset contained roughly $211,000$ events that were used to compute the required parameters. The remaining subset of logs for the period of September 2003 until April 2004 were used to check how accurate the MRP mobility predictor had performed for each user. The prediction results include both the next location transition and the time within which the transition occurred. It would have been more appropriate to construct separate MRP parameters for different periods in the academic term and the time-of-day. However, the limited size of the data set would not allow for a reasonably accurate fitting of the required parameters.

To assess the accuracy of the predictions, we simply chose to measure the number of times a correct prediction was made from the total number of prediction attempts per user. In other words, every one of the users' logs were processed for the events that led to a transition from either of the 21 access points to the set of possible neighboring locations, including the state where the session is terminated in the current location. Each of the events (per user) were then checked to see if the MRP mobility prediction and the one employing the conventional methods would have predicted the event correctly.

For checking the accuracy of the MRP mobility predictor, Equation (2.16) was

used to compute $Q_{i,k}(T \pm \Delta) = Q_{i,k}\left(\max\{0, T - \Delta\}, T + \Delta\right)$ for all possible future states $k \in \Omega_i$ from i, where Ω_i is the set of future states to which a user can transition from state i, T is the actual length of time a user spent in state i before making a transition, and Δ was chosen to be equal to 1 minute. The choice of the step-size Δ depends on how frequent the network needs to periodically predict the transition behavior of the user. In other words, Δ determines the time window size of when the computed prediction results are valid, after which a new prediction would need to be evaluated for the next time window if the user has not made any transition. Following a similar idea to the one proposed by Petzold et al. in [84], the accuracy of the proposed scheme was evaluated by simply checking to see if the model returned the correct prediction during the time window in which a transition occurred. Note that this accuracy is dependent on the time window Δ and the traffic data used to compute the model parameters.

For an event with an *actual* state transition of $i \rightarrow j$ and sojourn time of T units in state i, a correct prediction corresponds to having the result where $Q_{i,k}(T \pm \Delta)$ is a maximum for $k = j$. A similar approach was taken for the case of using transition probabilities alone. Hence, the average prediction accuracy metric Φ for each user can be computed as follows,

$$\Phi \;=\; \frac{\sum_{m=1}^{M} \varphi_{m,i,j}(T, \Delta)}{M}, \tag{3.1}$$

$$\text{where} \quad \varphi_{m,i,j}(T, \Delta) \;=\; \begin{cases} 1 & \text{if} \quad \max_{(k \in \Omega_i)} \left\{Q_{i,k}(T \pm \Delta)\right\} = Q_{i,j}(T \pm \Delta) \\ 0 & \text{otherwise} \end{cases},$$

and M is the number of relevant events in the user's log that were processed for computing Φ. This metric measures the prediction accuracy for spatial transitions

within a given time window. Notice that this result could also be described as a measure of *strict* accuracy since the predictions were checked to see if they were *strictly* correct or not. One could simply instead look at the percentage of accurate predictions across all the events tested and irrespective of the user. However, in this work it was chosen to measure the prediction accuracy from the perspective of each of the users in the data set rather than what is perceived by the network on the overall.

The performance of the proposed prediction scheme was compared with the conventional schemes that employ the location transition probabilities $P_{i,j}$ alone. Any of the predictors that relied on the usage of extra information that are not usually available from the traffic traces, e.g. geographical constraints, channel conditions, and user preferences, have been avoided in the comparison. Another common prediction scheme that has been considered by various researchers (e.g. [66]) is one that weights each of the transition probabilities with the corresponding mean sojourn times $\widetilde{\tau}_{i,j}$ such that,

$$\widetilde{P}_{i,j} \;=\; \frac{P_{i,j}\;\widetilde{\tau}_{i,j}}{\sum_k P_{i,k}\;\widetilde{\tau}_{i,k}}. \tag{3.2}$$

As mentioned earlier, some of the mobility models that had been previously developed by other researchers had assumed a sojourn time distribution that exhibit the "memoryless" property. In the previous analysis, the Log-normal distribution ($LogN(\mu, \sigma)$) was found to be the best fitting distribution, based on the fitting results achieved using Matlab, for the sojourn time distribution $G_{i,j}(t)$ from the data set. The performance of the proposed predictor was also compared with $G_{i,j}(t)$ assuming an exponential distribution ($Exp(\lambda)$) to test how well the distribution with the memoryless property can approximate the sojourn time behaviors in the predictions. The empirical distribution ($EmpD$) of the sojourn times in the data

set was also computed for $G_{i,j}(t)$ in the proposed predictor and its performance was compared with the others.

Table 3.1 shows a summary of the average prediction accuracy that the user would experience and for the transitions that are made from each of the 21 access points within the Library Building 2. The transitions include the event of a user terminating of the network session at the given access point. The accuracy of the proposed predictor was measured for each of the users in the data set using Equation 3.1 and the average across all the users are shown in the table. In general, the results do show that the proposed predictor has in some instances improved on the accuracy of the predictions that are returned using the transition probabilities alone. The "$\widetilde{P}_{i,j}$ Predictor" appears to show a stronger performance on the overall but such predictions do not include the time window at which the next transition might occur.

Amongst the proposed "$Q_{i,j}$ Predictor" with different $G_{i,j}(t)$ distributions, the results generally show that in most cases a better accuracy is achieved when the empirical distribution is used for $G_{i,j}(t)$. Furthermore, the Log-normal distribution appears to be a reasonable approximation of the empirical distribution, given by the accuracy results of the proposed predictor with $G_{i,j}(t) \sim LogN(\mu, \sigma)$ relative to $G_{i,j}(t) \sim EmpD$. Another important observation from the results is that the proposed predictor with $G_{i,j}(t) \sim Exp(\lambda)$ has in a lot of cases shown a lower performance relative to the others. This illustrates one of the advantages of the proposed mobility model not having its sojourn time behavior restricted to any particular distribution. The bar graph in Figure 3.9 illustrates the difference in the performance of the proposed "$Q_{i,j}$ Predictor" with the three different distributions $G_{i,j}(t)$.

The results in Table 3.1 are for the case where the $Q_{i,j}$ Predictor was applied with $\Delta = 1$. The time window at which these predictions were also varied by increasing

Tab. 3.1: Summary of the overall prediction accuracy results for single transitions from the access points in Library Building 2, with $\Delta = 1$.

Transitions From AP	Conventional Predictors		Proposed Predictor Using MRP			Total # of Transitions	Total # of Users
	$P_{i,j}$ Predictor	$\widetilde{P}_{i,j}$ Predictor	$Q_{i,j}(t)$ Predictor				
			$G_{i,j}(t) \sim LogN(\mu, \sigma)$	$G_{i,j}(t) \sim Exp(\lambda)$	$G_{i,j}(t) \sim EmpD$		
AP1	0.68271	0.68271	0.54818	0.45476	0.54996	9654	2335
AP2	0.65925	0.65925	0.45632	0.41325	0.55478	1583	373
AP3	0.21646	0.58399	0.46598	0.4313	0.47495	17663	2709
AP4	0.61665	0.61665	0.48976	0.37077	0.48891	10467	1551
AP5	0.74455	0.74455	0.74455	0.59785	0.74872	6180	1240
AP6	0.68193	0.68193	0.59172	0.51833	0.61324	9861	1790
AP7	0.74285	0.74285	0.62492	0.54076	0.64831	17654	902
AP8	0.63339	0.63339	0.53101	0.498	0.56918	1950	551
AP9	0.8422	0.8422	0.80873	0.8422	0.8422	18065	1959
AP10	0.10901	0.47234	0.28694	0.29196	0.34441	13425	2424
AP11	0.84718	0.84718	0.84718	0.84718	0.84718	21025	2396
AP12	0.59599	0.59599	0.59599	0.59599	0.59601	3405	329
AP13	0.6673	0.6673	0.6673	0.56802	0.6673	9332	1612
AP14	0.19765	0.5822	0.50558	0.46924	0.52539	42481	3075
AP15	0.61539	0.61539	0.61539	0.61539	0.61542	13720	208
AP16	0.076945	0.64095	0.26693	0.25402	0.25206	11694	1409
AP17	0.20611	0.56278	0.40093	0.40455	0.40407	34996	3043
AP18	0.093849	0.43435	0.22364	0.20206	0.37364	1153	137
AP19	0.6665	0.6665	0.57832	0.46232	0.57832	22672	3106
AP20	0.45281	0.45281	0.54859	0.5705	0.54668	3355	880
AP21	0.69175	0.69175	0.57861	0.4857	0.60016	26732	3487
Overall	0.525736857	0.638907619	0.541741429	0.496864286	0.563851905		

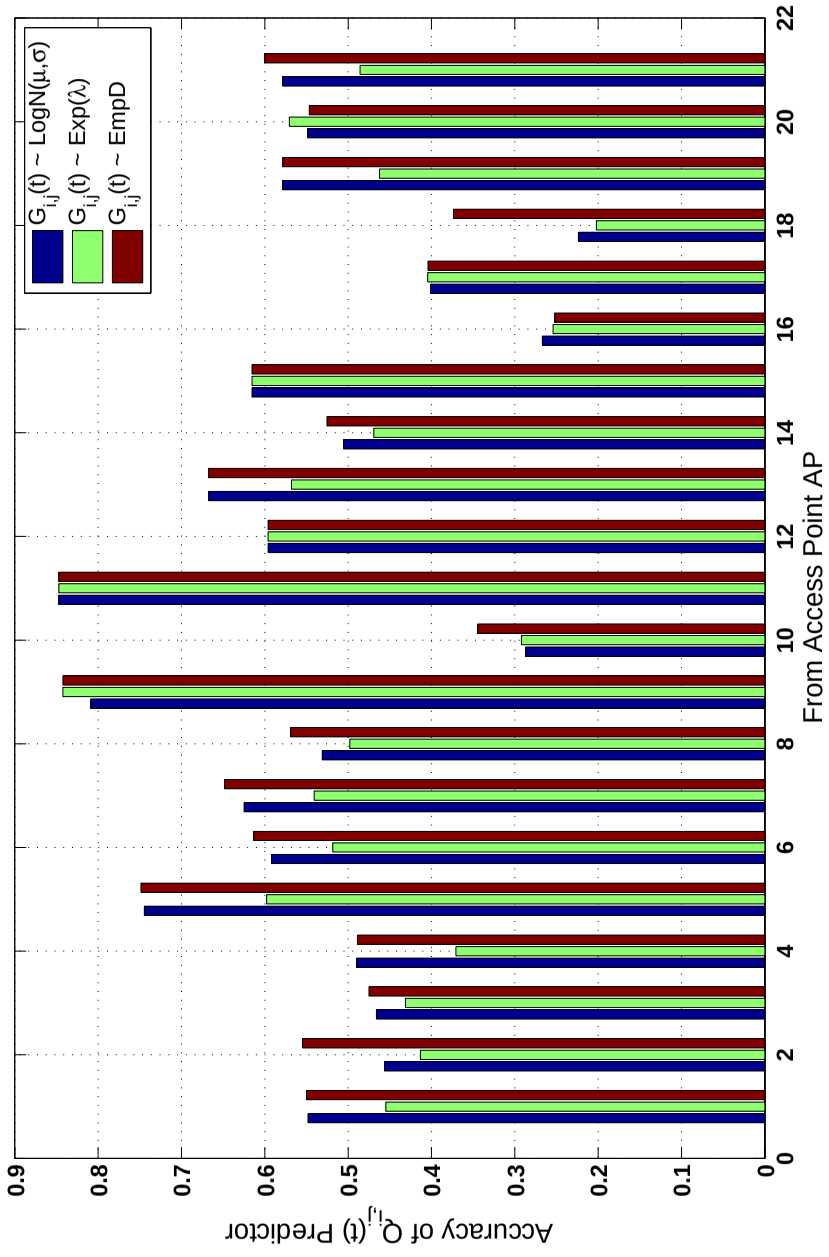


Fig. 3.9: The performance of the proposed "$Q_{i,j}$ Predictor" with the three different distributions $G_{i,j}(t)$.

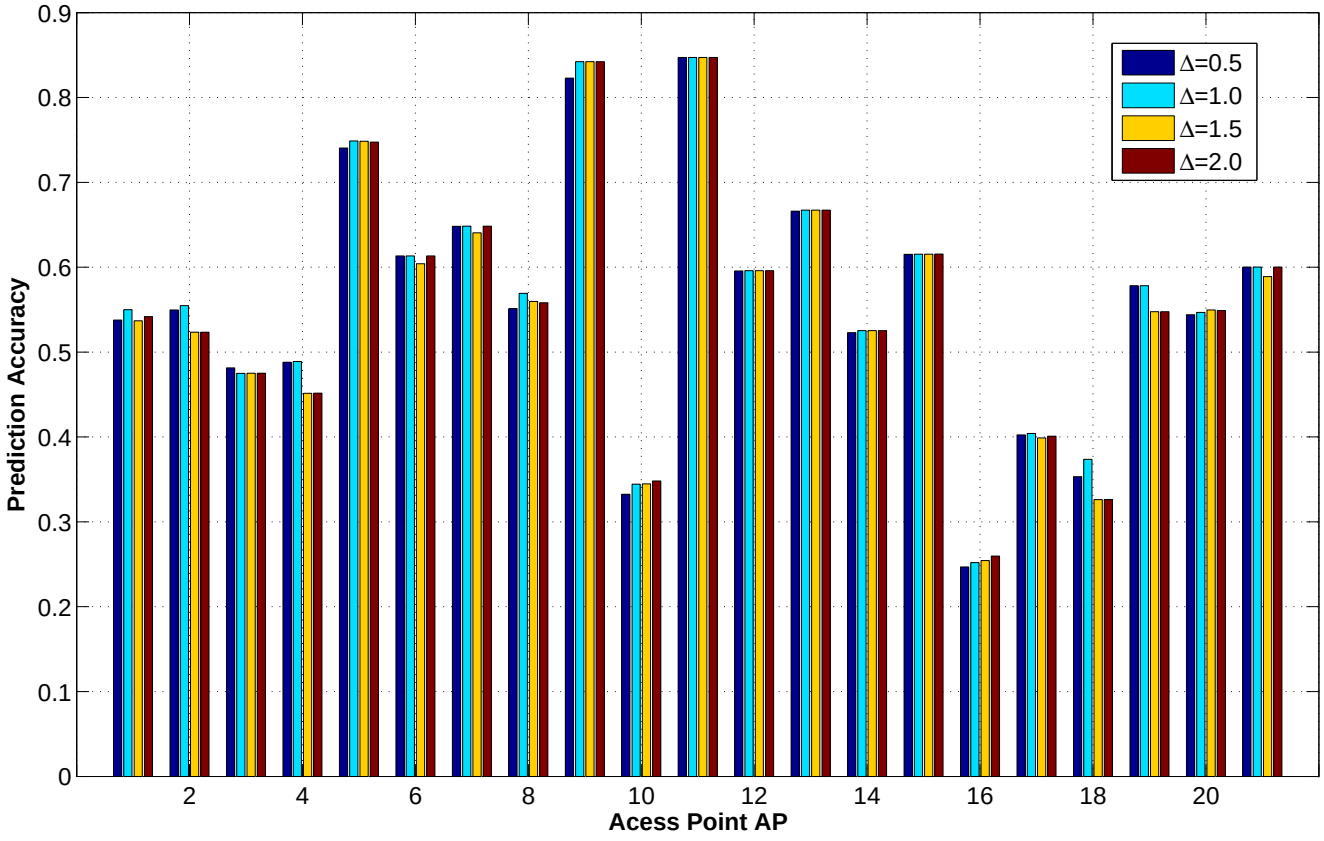
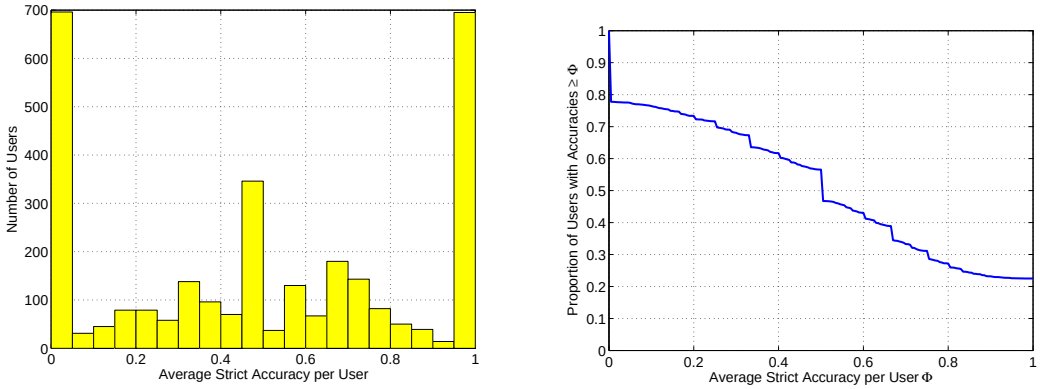


Fig. 3.10: The performance of the proposed "$Q_{i,j}$ Predictor" with $G_{i,j}(t) \sim EmpD$ and under various choices of Δ.

the value of Δ. As an example, Δ was varied from 0.5 to 2 minutes. This change in Δ had an insignificant effect on the accuracy results for $G_{i,j}(t) \sim LogN(\mu, \sigma)$ and $G_{i,j}(t) \sim Exp(\lambda)$. However, the changes in Δ did have an influence on the prediction accuracy results when applying the $Q_{i,j}$ predictor with $G_{i,j}(t) \sim EmpD$ and the difference in performance is shown in Figure 3.10. The results show that an optimum value of Δ for generating the predictions from each of the 21 access points are not the same. The assignment of $\Delta = 1$ appears to be the better choice in most cases.

To help better understand the results, a more thorough examination of the prediction accuracy results will be given and focusing on the cases of transitions made from AP14 and AP17. Figures 3.11 and 3.12 show the results of the average prediction accuracies after evaluating the data using both the proposed $Q_{i,j}(t)$predictor with $G_{i,j}(t) \sim LogN(\mu, \sigma)$ and the transition probabilities $P_{i,j}$ alone, respectively. These results were for the case of users transitioning away from the location served by AP14. Part (a) in each figure shows a histogram of the number of users in the data set that exhibited a certain range of average prediction accuracies Φ. Part (b) provides a cumulative plot of the same results for the proportion of users with average prediction accuracies Φ above a particular level. Each of the logs in the data set were scanned for the relevant events and the occurrences were checked to see if the $Q_{i,j}(t)$ predictor would have predicted them accurately based on Equation (3.1). In the case of the results given in Figures 3.11 and 3.12, these relevant events are those that involve a transition from AP14 to a neighboring access point within the same building.

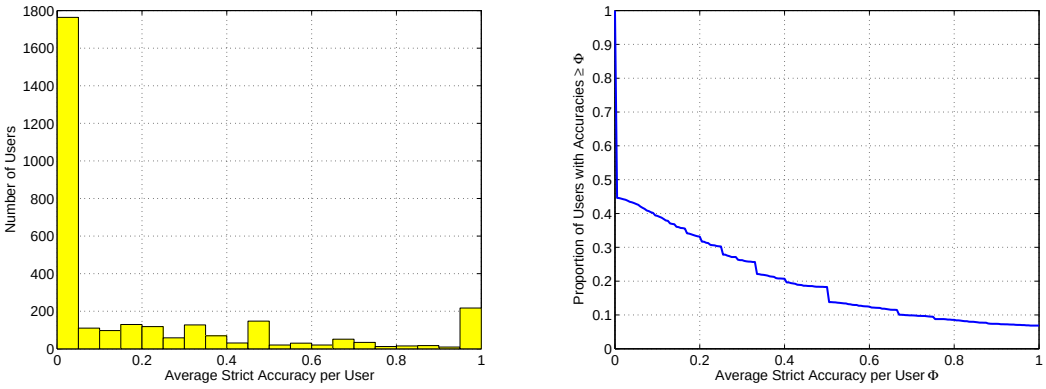
The results in Figure 3.11 suggest that the predictions made using the proposed $Q_{i,j}(t)$ predictor are an improvement on those returned using the transition probabilities alone, based on the results shown in Figure 3.12. A better overall accuracy



(a) Histogram.

(b) Fraction of users with Φ above a certain level.

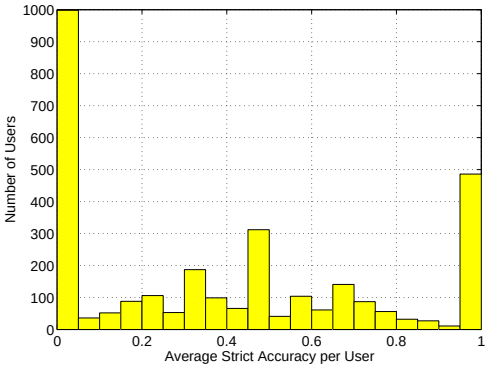
Fig. 3.11: Average prediction accuracies Φ for transitions made from AP14 using the proposed MRP model.



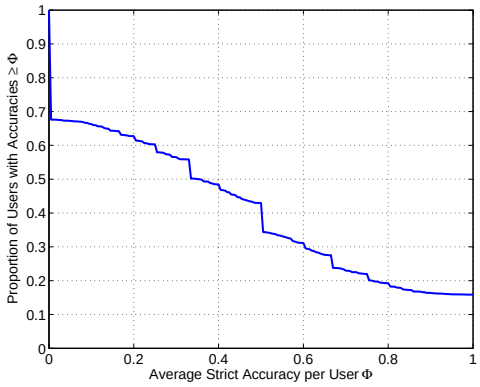
(a) Histogram.

(b) Fraction of users with Φ above a certain level.

Fig. 3.12: Average prediction accuracies Φ for transitions made from AP14 using transition probabilities $P_{i,j}$ only.

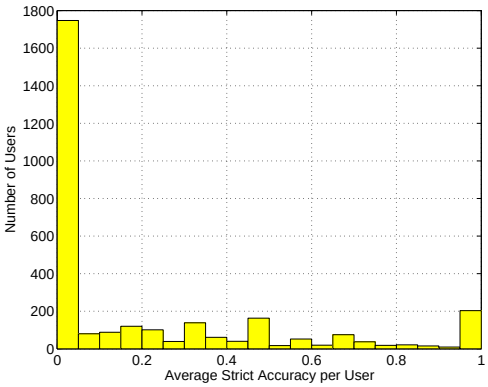


(a) Histogram.

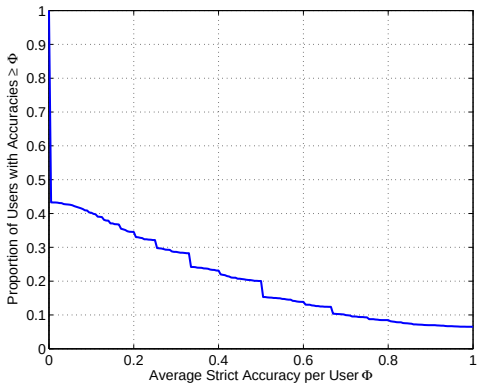


(b) Fraction of users with Φ above a certain level.

Fig. 3.13: Average prediction accuracies Φ for transitions made from AP17 using the proposed MRP model.



(a) Histogram.



(b) Fraction of users with Φ above a certain level.

Fig. 3.14: Average prediction accuracies Φ for transitions made from AP17 using transition probabilities $P_{i,j}$ only.

is reflected by the heavier tail in the plot shown in Figure 3.11(b). The same conclusion can be given for users that are transitioning out of AP17 when comparing the results in Figures 3.13 and 3.14. These second set of results also show that higher accuracies have been achieved when predicting the next transition from AP17 using the proposed mobility model, when compared with using the transition probabilities alone, as given by the results in Figures 3.13 and 3.14, respectively. In general, the positive outcome of the results collected from applying the $Q_{i,j}(t)$ predictor also indicate a possible existence of a relationship between the user's mobility and the location sojourn times (or the times the session is associated with an access point).

The results in Figures 3.11 to 3.14 also reveal that there had been many instances where the prediction accuracy is close to 0%. Each of the scanned user logs did not contain the same number of events to be processed which may have had an adverse effect on the overall results. For instance, out of the nearly 1000 logs that had close to 0% prediction accuracies from AP14, most of these logs contained a very small number of relevant events for processing. Figure 3.15 show the number of users that had between 1 and 50 transitions from AP14 in the subset of the data that was processed for the predictions. The figure shows how many of those users had a very low number of events and in most of these cases the predictions were unsuccessful. There were a few users which had a much higher number of events with the maximum being 2118. Figure 3.16 show the number of users that had between 1 and 50 transitions from AP17 and also shows a larger number of users with a very low number of events. These results seem to suggest that users in WLAN environments tend to be less mobile which could have had some influence on the prediction results. Nevertheless, the proposed $Q_{i,j}(t)$ predictor displayed an improvement on the overall with a higher prediction accuracy than what can be achieved using the transition probabilities alone.

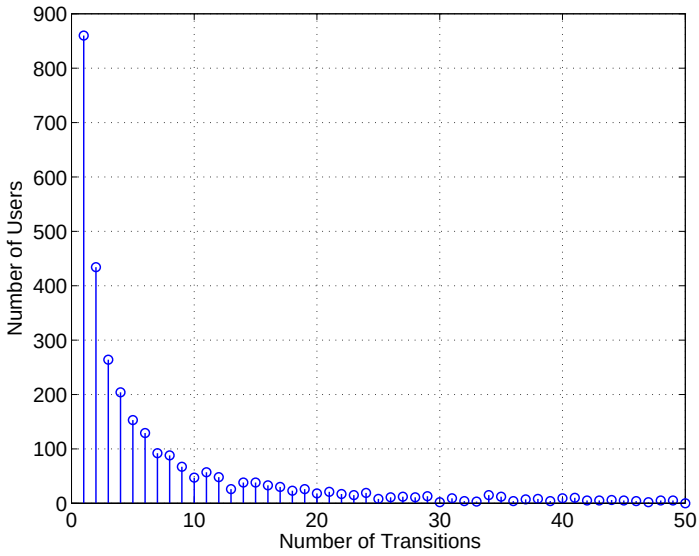


Fig. 3.15: The number of users that had between 1 and 50 transitions from AP14.

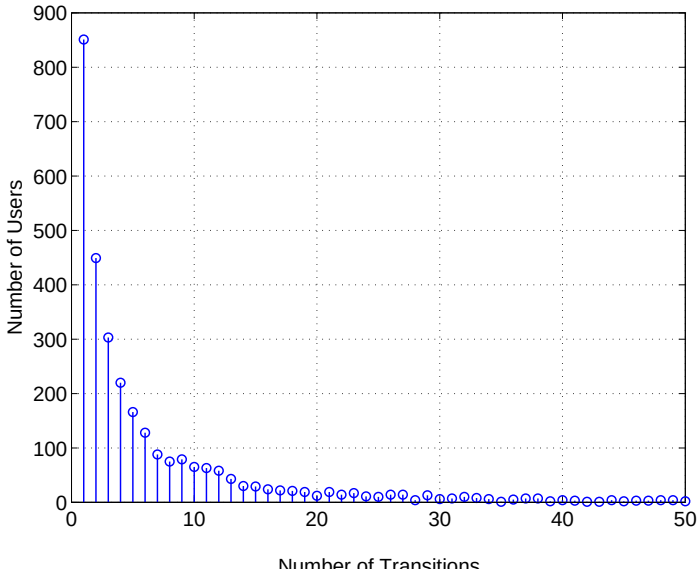


Fig. 3.16: The number of users that had between 1 and 50 transitions from AP17.

In the case of transitions from AP14, around 50.6% of the next transitions were accurately predicted using the proposed $Q_{i,j}(t)$ predictor, compared to the 19.8% accuracy using the transition probabilities $P_{i,j}$ alone. This would be the average accuracy perceived by each user. Thus, the proposed model was able to improve on the predictions by roughly 156% in this example. For the second case of transitions from AP17, an overall prediction accuracy of 40.1% with the proposed $Q_{i,j}(t)$ predictor was observed and improved on the predictions using $P_{i,j}$ alone by around 95%.

Order-n Markov predictors have also been shown to improve on the accuracy of the predictions [27–29]. The analysis was extended to examine the performance of the proposed predictor using the order-2 semi Markov kernel $Q_{h,i,j}(t)$ (see Equation (2.15)). The data set was processed again for computing the elements $P_{h,i,j}$ and $G_{h,i,j,}(t)$ that were needed to construct the proposed predictor. The performance of the $Q_{h,i,j}(t)$ predictor was compared with the conventional predictors that employ the transition probabilities $P_{h,i,j}$ alone. Such predictors consider both the current state and the previous state of the user. The performance of the "$\widetilde{P}_{h,i,j}$ Predictor" was also included in the analysis and can be computed in a manner that is similar to the one given in Equation (3.2).

Table 3.2 summarizes the prediction accuracy results gained by apply the order-2 semi-Markov $Q_{h,i,j}(t)$ predictor in contrast with the conventional predictors that employ the location transition probabilities $P_{h,i,j}$ alone. Note that the total number of transitions are not the same as those in Table 3.1 since each user's log must contain at least 2 transitions in order to apply the order-2 Markov and semi-Markov predictors. The results are for the case of single transitions from AP14 and AP17. In both cases, the accuracy of the $Q_{h,i,j}(t)$ predictor was found to be higher than what was achieved when using both the $P_{h,i,j}$ and the $\widetilde{P}_{h,i,j}$ predictor. However, the

Tab. 3.2: Summary of the overall prediction accuracy results for single transitions from AP14 and AP17 in Library Building 2, using order-2 Markov predictors and with $\Delta = 1$.

	Transitions From AP14	Transitions From AP17
$P_{h,i,j}$ Predictor	0.36689	0.36205
$\widetilde{P}_{h,i,j}$ Predictor	0.36694	0.42715
$Q_{h,i,j}(t)$ Predictor with $G_{h,i,j}(t) \sim LogN(\mu, \sigma)$	0.45481	0.45041
$Q_{h,i,j}(t)$ Predictor with $G_{h,i,j}(t) \sim Exp(\lambda)$	0.5005	0.72949
$Q_{h,i,j}(t)$ Predictor with $G_{h,i,j}(t) \sim EmpD$	0.45774	0.4593
Total Number of Transitions	21072	19789
Total Number of Users	1913	1996

more interesting result is the accuracy gained by using the $Q_{h,i,j}(t)$ predictor with $G_{h,i,j}(t)$ being approximated by the exponential distribution with the memoryless property. Even though the Log-normal distribution was found to be a better fit for describing the behavior of the sojourn times, the results suggest that the exponential distribution might be a better approximation in describing the sojourn time behaviors for the case of order-2 semi-Markov predictors.

Figure 3.17 compares the prediction accuracies achieved using both the order-1 and order-2 Markov and semi-Markov predictors and for single transitions made from AP14 and AP17. In the case of transitions from AP14, the order-2 semi-Markov $Q_{h,i,j}(t)$ predictor had not generally improved on the order-1 semi-Markov $Q_{i,j}(t)$ predictor. A slight improvement was observed if the behavior of the sojourn times was assumed to follow an exponential distribution. The case of transitions from AP17 shows a different performance whereby the order-2 semi-Markov $Q_{h,i,j}(t)$ predictor had exhibited some improvement over the order-1 semi-Markov $Q_{i,j}(t)$ predictor, especially when $G_{h,i,j}(t) \sim Exp(\lambda)$. Furthermore, $Q_{i,j}(t)$ predictor in this

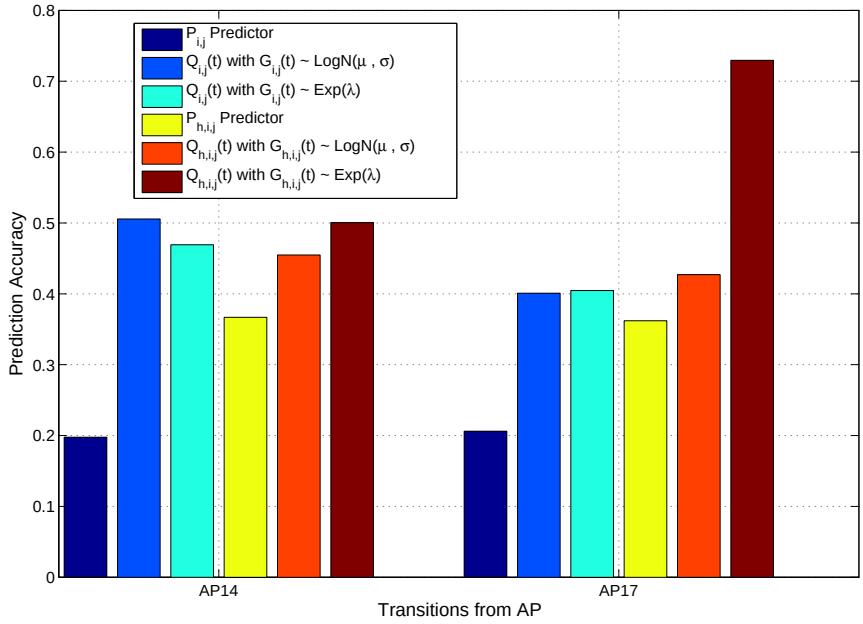


Fig. 3.17: The difference in performance between the order-1 and order-2 Markov and semi-Markov predictors, for transitions from AP14 and AP17.

example was successful at gaining a higher accuracy without the added knowledge of the previous state of the user that was needed for the $P_{h,i,j}$ predictor.

The assignment of Δ was also varied to examine the influence it might have on the performance of the $Q_{h,i,j}$ predictor. Similar to the previous example, varying Δ from 0.5 to 2 minutes had changed the accuracy by less than 1% for the cases where $G_{h,i,j}(t) \sim LogN(\mu, \sigma)$ and $G_{h,i,j}(t) \sim Exp(\lambda)$. With $G_{h,i,j}(t) \sim EmpD$, the changes in the prediction accuracy results were quite noticeable and are shown in Figure 3.18. In the case of making predictions from AP14 and AP17, the results show that the choice of $\Delta = 1$ has yielded the better performance in terms of the prediction accuracy.

To improve on the prediction accuracies, one way would be to consider the temporal period/season at which the predictions are to be made. For example,

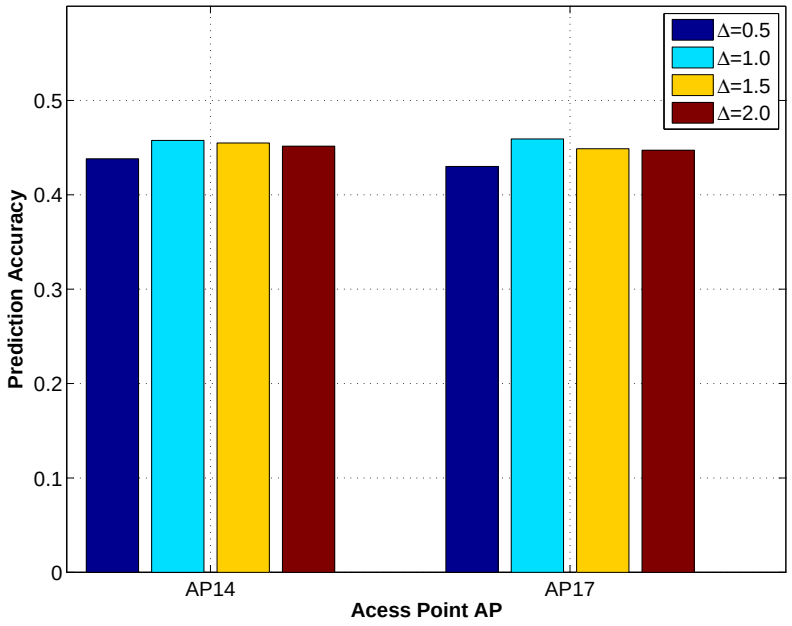


Fig. 3.18: The performance of the proposed "$Q_{h,i,j}$ Predictor" with $G_{h,i,j}(t) \sim EmpD$ and under various choices of Δ.

the behavior of users during the day time may differ from those during the night time. Hence, one possible approach would be to have two different semi-Markov mobility models for making predictions during those different times of the day. This approach can be taken further to consider the different days of the week as well as the different months of the year. While this proposition will likely lead to a more accurate model for the mobility behavior of the users, it does require an extensive amount of mobility history to be processed.

3.3 Location Approximation

With the prediction scheme that has been proposed so far, a crucial matter yet remains to be addressed which may be the cause of some concern to network managers and operators. The amount of data needed to construct the kernels in the MRP model can be quite large, especially when a vast number of elements are required to be defined for a given network. If we take the example of the 7-cell structure shown in Figure 2.1 as being the entire network coverage area under consideration, then the number of elements that need to be defined at most is 62, as given by the number of non-zero elements in Equation (2.10). These elements capture the time-varying probabilities of a user being mobile between certain neighboring cells amongst the 7 distinct locations while undergoing either an active or idle session with the network, as well as the transitions involving the changes in the session activity. An MRP-based mobility model for such a network involves transitions between any of the 14 possible states of the system defined by the following state space $\mathcal{X} = \{(-1, -2, \cdots, -7)\,; (1, 2, \cdots, 7)\}$. The details for such a model are shown in the block matrices given by Equations (2.10) to (2.14).

The amount of processing needed to evaluate the necessary predictions may be altogether laborious, depending on the number of different elements involved in the computation. The size of the MRP-based mobility model's state space could have a direct influence on the amount of processing involved with generating the predictions, since it would involve querying a central database that contain the details of the set of semi-Markov elements. In Section 2.3, it was mentioned that one possible approach for reducing the size of the state space would be to restrict the model's focus to the mobility behavior of users with active sessions alone, thereby ignoring the state of the users when they are idle. This is similar to what was accomplished

by the numerical analysis examined throughout this chapter. However, this sort of reduction in state space might not be sufficient, especially when dealing with networks with many locations. Another approach to further scale down the state space of the mobility model would be to *appropriately* cluster the entire set of locations into certain groups. Each cluster would include a group of neighboring locations that are individually served by their own wireless access point. The clustering assists with reducing the model's state space and subsequently relieving some of the processing power needed for computing the predictions.

The paucity of traffic data may also be a good reason to apply the location clustering. The grouping of the location transition behaviors has the effect of combining together the traffic data recorded at each of the locations in the cluster into a bigger set. This new set of data could improve on the precision in computing the semi-Markov elements $Q_{i,j}(t)$, especially the fitting of the sojourn time distributions $G_{i,j}(t)$.

The clustering approach tends to approximate the mobility pattern details of a user in the network. Consequently, a user in such a case is said to be within the service coverage of a group of access points that belong to the same cluster, without identifying the exact access point that is serving the mobile user. However, as with most approximations, this location clustering approach could reduce the accuracy and/or precision of the prediction computations. From the network manager's perspective, the approximation may come at the cost of having the predictions being less explicit. This could also create the need for future resource reservations at multiple locations within the single cluster, rather than confining the resource reservation decision to a sole location. Moreover, while it may matter less in some cases, the choice of how these various locations are assigned to a certain cluster could have a significant impact on the performance of the semi-Markov predictor.

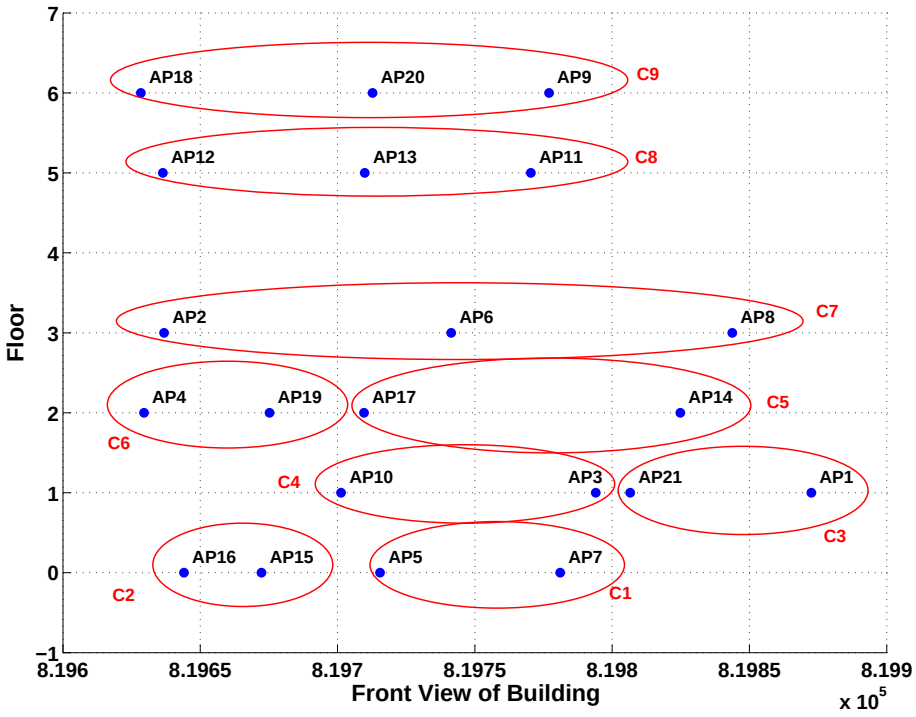


Fig. 3.19: A front view of the spatial locations of the 21 wireless access points in Library Building 2 at Dartmouth College, with clustering.

The traffic traces provided by [83] also supplied the details of where the wireless access points are deployed across the Dartmouth College campus. The information included both the coordinates and the floor number of where these access points are located in the Library Building 2 which was collected from the AutoCAD drawings of the premises. An outline of where these 21 access points are located in the building is shown in Figure 3.19, as viewed from the front side of the Library Building 2. The figure also shows how these access points are distributed across the various floors of the building.

To demonstrate the usage and performance of including the AP location clustering into the semi-Markov prediction scheme, the same set of traffic traces from [83] were used again which included the mobility behavior of users between the access

points in Library Building 2. The clustering of those locations were chosen to be intuitively assigned as given in Figure 3.19. The decision behind the assignment of these clusters relate to the traffic loads that were observed in the traces. For instance, it was noticed that the majority of the mobility traffic was concentrated on the lower 3 floors of the building while the remaining upper floors exhibit relatively less traffic. Hence, it was elected to have the access points in each of the upper floors clustered together as a single location while two clusters were assigned for the set of access points in the lower floors, as shown in Figure 3.19. The decision on how to optimally assign these location clusters would rely on the knowledge of other relevant factors which will be investigated in future works. In essence, the aim here is to allocate these clusters for the purpose of reducing the number of states that are needed to define the MRP-based predictor and examine how well it performs.

The elements for the $Q_{i,j}(t)$ predictor utilizing the cluster location approximation were constructed in a manner that is similar to the one adopted earlier in this chapter. The set of traffic traces that had their logs time-stamped prior to the year 2003 were processed to compute the cluster location transition probabilities $P_{i,j}$, as well as the conditional sojourn time distribution $G_{i,j}(t)$, for evaluating the $Q_{i,j}(t)$ elements defined by Equation (2.1). In this approximation, a transition due to mobility is caused by a change in clusters rather than access points. Thus, a user is said to remain in the same cluster location even if a transition is made between two access points that are located within the same cluster. The state-dependent sojourn time is the duration at which a user's network session remains active within the same cluster. In this particular case, the set of possible states for this MRP-based predictor includes the 9 states that signify the set of all possible cluster locations that a user can transition between. In following the same idea of the original semi-Markov predictor, two additional states were also defined. The first additional state

being the OFF state at which a user terminates its session in the same cluster location, along with an $OTHER$ state that describes the location of an access point outside the Library Building 2 (which can also be thought of as another cluster of access points). Hence, in total this approximation has reduced the original 23 state semi-Markov predictor to one with 11 states. This scale of state space reduction is dependent on how one chooses to assign the clusters.

After processing the traffic traces for the $Q_{i,j}(t)$ elements, the predictor was then applied on the remaining set of traffic traces that were logged from the year 2003 onwards to examine the accuracy of the predictions with the applied approximation. Each user's sequence of transition logs were scanned to verify if the next transition state, as well as the time window, would have been correctly predicted according to Equation (2.17). The time window chosen was $\Delta = 1$. Each user's average prediction accuracy was examined using Equation (3.1) for single transitions from the 9 clusters. The overall average accuracy results from the $Q_{i,j}(t)$ predictor with location approximation was compared with predictors that only utilizes the transition probabilities $P_{i,j}$ and $\widetilde{P}_{i,j}$ from Equation (3.2). The distribution for the sojourn times were fitted as a Log-normal distribution, i.e. $G_{i,j}(t) \sim LogN(\mu, \sigma)$. Table 3.3 shows a summary of the prediction accuracy results for users transitioning from the 9 cluster locations. The results an overall lower prediction accuracy when applying the proposed $Q_{i,j}(t)$ predictor as compared with accuracy achieved by using the $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictor.

One of the clusters was chosen for a more detailed analysis of its prediction accuracy results, namely cluster $C5$. Note that $C5$ groups together the locations served by AP14 and AP17. Table 3.4 summarizes the prediction accuracies averaged over all the users in the entire set of traffic traces from [83], for transitions from cluster C5. The results in the first two columns were those found in the previous section

Tab. 3.3: Summary of prediction accuracy results for transitions from all 9 cluster locations.

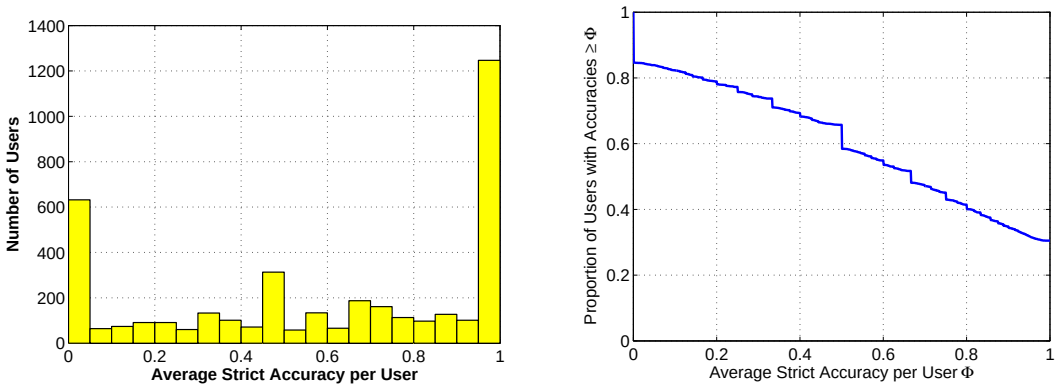
	$P_{i,j}$ Predictor	$\widetilde{P}_{i,j}$ Predictor	$Q_{i,j}(t)$ Predictor	No. of Transitions
From $C1$	0.694599	0.694599	0.548300	24295
From $C2$	0.601741	0.601741	0.499325	24401
From $C3$	0.645761	0.645761	0.604273	37516
From $C4$	0.504790	0.504790	0.451918	32950
From $C5$	0.600378	0.600378	0.575074	69953
From $C6$	0.646438	0.646438	0.646438	34012
From $C7$	0.652621	0.652621	0.586074	14188
From $C8$	0.782934	0.782934	0.782934	32955
From $C9$	0.815634	0.815634	0.815634	18995
Overall	0.661	0.661	0.6122	

without applying the location approximation (see Table 3.1) and are included in this table for comparison. The third column summarizes the overall average prediction accuracies using the conventional methods (i.e. the $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictors) and the $Q_{i,j}(t)$ predictor with location approximation. For each of the three types of predictors, the use of the location clustering approximation has improved on the prediction accuracies reported in the previous section. For example, the $Q_{i,j}(t)$ predictor was successful at accurately determining 57.5% of the transitions from cluster $C5$ which includes all users associated with both AP14 and AP17. These predictions did not distinguish the access point in $C5$ that was involved in the transition, which would have otherwise been achieved with an accuracy of 50.6% and 40.1% when applying the $Q_{i,j}(t)$ predictor without the location approximation for transitions from AP14 and AP17, respectively.

The results in the third column of Table 3.4 also reveal that in applying the

Tab. 3.4: Comparison of prediction accuracy results for transitions from AP14, AP17 and C5.

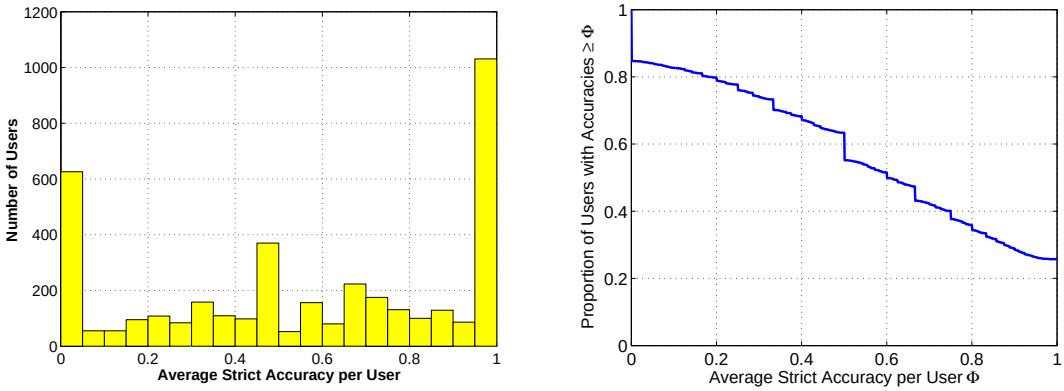
	From AP14	From AP17	From $C5$
$P_{i,j}$ Predictor	0.19765	0.20611	0.600378
$\widetilde{P}_{i,j}$ Predictor	0.5822	0.56278	0.600378
$Q_{i,j}(t)$ Predictor	0.50558	0.40093	0.575074



(a) Histogram.

(b) Fraction of users with Φ above a certain level.

Fig. 3.20: Average prediction accuracies for transitions made from C5 using the transition probabilities $P_{i,j}$ alone.



(a) Histogram.

(b) Fraction of users with Φ above a certain level.

Fig. 3.21: Average prediction accuracies for transitions made from C5 using the MRP-based model.

approximation, the $Q_{i,j}(t)$ predictor performed not as well as the location approximated $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictors. However, the accuracy results were fairly close and the $Q_{i,j}(t)$ included temporal estimates whereas the other two schemes were limited to predicting the next location transitions alone. Figures 3.20 and 3.21 illustrate the distribution details of the average prediction accuracy results Φ across all the users that were examined in the traffic data, after applying the location approximated $P_{i,j}$ and $Q_{i,j}(t)$ predictor, respectively. The results altogether show how these two predictors with the applied location approximation had a similar performance in terms of accurately predicting the next transitions. Despite the marginally low difference in performance, the $Q_{i,j}(t)$ predictor has the added capability of including temporal information in its future state estimations. This may render it a more favorable technique amongst the network managers and for yielding more informative predictions.

For transitions from all the 9 clusters, the $Q_{i,j}(t)$ predictor with location approximation had a consistently lower performance when compared with the other two predictors. This consistency necessitated a closer look at the results to identify factors that may have had an influence on the performance. A thorough scan of the results unveiled that the majority of the predictions were unsuccessful with determining the termination of a user's session in the current cluster, i.e. transitions to the OFF state. This might only be true for the data that was used in the analysis and it was difficult to confirm this conclusion without investigating other independent sets of traffic data in parallel. Another conceivable reason could be due to the method chosen for the assignment of the clusters.

To explore whether predicting the transitions to the OFF state had a negative impact on the performance of the location approximated $Q_{i,j}(t)$ predictor, the same numerical analysis was repeated while ignoring the transitions to the "OFF" state,

in both in the model and the prediction evaluations. In other words, this analysis only covered the cluster location transitions of the users with active sessions, thus reducing the size of the state space in the MRP predictor to 10. The results in Table 3.5 provide the overall prediction accuracies reported using the location approximated $Q_{i,j}(t)$ predictor and comparing them with those returned from employing the $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictors. In general, even though the values may seem lower than what was achieved earlier, the results now show that the location approximated $Q_{i,j}(t)$ predictor has in most cases displayed a better performance than the other two predictors. For example, the location approximated $Q_{i,j}(t)$ predictor was found to have accurately predicted 63.7% of the transitions from cluster $C9$, which was better than the 55.9% accuracy achieved by using the $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictors.

In summary, including the cluster location approximation into the proposed MRP-based mobility prediction can assist in reducing the size of the state space and the amount of computations needed to construct the model. The numerical

Tab. 3.5: Summary of prediction accuracy results for location transitions alone from all 9 cluster locations, without predicting the "OFF" state.

	$P_{i,j}$ Predictor	$\widetilde{P}_{i,j}$ Predictor	$Q_{i,j}(t)$ Predictor	No. of Transitions
From $C1$	0.403056	0.403056	0.433599	11134
From $C2$	0.385997	0.385997	0.385997	10847
From $C3$	0.413190	0.362497	0.499675	14627
From $C4$	0.413150	0.207022	0.417124	19352
From $C5$	0.382102	0.382102	0.382102	32830
From $C6$	0.381496	0.246572	0.405470	12970
From $C7$	0.389580	0.264418	0.441683	5888
From $C8$	0.592096	0.592096	0.592096	5454
From $C9$	0.559315	0.559315	0.637209	2360
Overall	0.436	0.378	0.466	

examples have shown that the location approximation was successful at achieving an adequate level of performance when compared with the conventional $P_{i,j}$ and $\widetilde{P}_{i,j}$ predictors. However, this comes at the cost of reducing the spatial details needed for future resource reservation purposes. Nevertheless, the results reported in this section do illustrate the promising potential in applying the cluster location approximation in conjunction with the proposed MRP-based predictor, especially when a reduction in the size of the state space in the proposed model is favorable.

4. FURTHER APPLICATIONS OF THE SEMI-MARKOV MOBILITY PREDICTION

4.1 Multi-Transition Mobility Prediction

This section will show how to extend the kernel definitions given in Equations (2.1) and (2.6) to predict the next Nth transition, as well as estimating the time-varying probability of finding a user at a particular state after N transitions. For simplicity, a discrete-time semi-Markov model will be assumed throughout this chapter and the continuous-time case can be developed in a similar manner. This assumption allowed for an easier presentation of the results. The state sojourn times are considered in discrete-time. The discrete-time distributions $G_{i,j}(t)$ (and subsequently $g_{i,j}(t)$) will be chosen such that $q_{i,j}(0) = Q_{i,j}(0) = 0$. This implies that the sojourn time in each state must be at least 1 unit of time. Furthermore, it is assumed that each transition requires at least 1 time unit and that no more than a single transition in each state is permitted during any unit of time. Hence, the property $t \geq N$ is always true, which implies that the minimum time for making N transitions is $t = N$.

In general, the kernel of the Nth transition prediction can be defined as follows,

$$q_{i,j}^{N}(t) \;=\; Pr\left\{X_{n+N} = j \;\; T_{n+N} - T_n = t \mid X_n = i\right\}, \tag{4.1}$$

$$\text{with} \quad Q_{i,j}^{N}(t) \;=\; \sum_{\tau=N}^{t} q_{i,j}^{N}(\tau), \tag{4.2}$$

for $t \geq N$. The elements $Q_{i,j}^N(t)$ denote the probability that immediately after making the transition into state i, the user is in state j by the Nth transition and in an amount of time less than or equal to t. For example, if the states are the cell IDs, then $Q_{1,4}^2(7)$ is the probability that the user makes a transition from location 1 to another location and then followed by a transition into location 4, all within 7 units of time from entering location 1.

The semi-Markov transition matrices $\mathbf{Q}(t) = \{Q_{i,j}(t)\}$ and $\mathbf{q}(t) = \{q_{i,j}(t)\}$ can be constructed using the semi-Markov kernels defined in Equations (2.1) and (2.6), respectively. Hence, the matrix $\mathbf{q}^N(t) = \left\{q_{i,j}^N(t)\right\}$ can be computed recursively as follows,

$$\mathbf{q}^N(t) \;=\; \sum_{\tau=N-1}^{t-1} \mathbf{q}^{N-1}(\tau)\mathbf{q}(t-\tau), \qquad \text{for} \quad t \geq N \geq 1, \tag{4.3}$$

$$\text{where} \quad q^0(t) \;=\; \begin{cases} 0 \;\;, & t > 0 \\ 1 \;\;, & t = 0 \end{cases} \;\;,$$

$$\text{with} \quad \mathbf{Q}^N(t) \;=\; \sum_{\tau=N}^{t} \mathbf{q}^N(\tau). \tag{4.4}$$

The elements defined in Equations (4.1) and (4.2) can be derived from the matrices $\mathbf{q}^N(t)$ and $\mathbf{Q}^N(t)$, respectively.

Using the expression for $\mathbf{Q}^N(t)$, we can further define the following,

$$\mathbf{\Lambda}^N(t) \;=\; \sum_{n=1}^{N} \mathbf{Q}^n(t) \qquad \text{for} \quad t \geq N, \tag{4.5}$$

where the element $\Lambda_{i,j}^N(t)$ is the probability that a transition from state i to state j occurs after N or fewer transitions and within time t from entering state i.

With the same traffic traces that were used in the previous chapter, namely from [83], we next show how one can apply and interpret the results gained from

employing the proposed multi-transition mobility predictions. The matrix $\mathbf{q}(t)$ was evaluated using the same transition probabilities and the sojourn time distributions that covered the transitions amongst the 21 access points in the Library Building 2. This had allowed for computing both $\mathbf{q}^N(t)$ and $\mathbf{Q}^N(t)$ using Equations (4.3) and (4.4), respectively. In the following numerical example, the transitions made from AP17 to AP1 were considered. According to the information supplied with the traffic data, this particular transition involved a user having to move from the second floor to the first.

The plots in Figures 4.1 and 4.2 show the probabilities of an arbitrary user with an active session at AP17 eventually having its session associated with AP1, after completing N transitions *at* and *within* t minutes of the session remaining active from AP17, respectively. Out of the 4 possibilities shown in Figures 4.1 and 4.2, the higher probability $Q_{17,1}(5)$ suggests that a user will most likely complete a direct and single transition from AP17 to AP1 if the transition occurred within about 5 minutes of the session being associated with AP17. In other words, if a transition were to occur from AP17 to AP1 and within 5 minutes of initially being in AP17, then it would most likely involve a single transition. This direct transition becomes the least likely event if the session remains active beyond 20 minutes, as shown by the plot for $q_{17,1}(t)$ in Figure 4.1. For sessions that remain active beyond 5 minutes of being initially associated with AP17, a user is more likely to end up having associated its session with AP1 after making a transition to another access point from AP17 before reaching AP1, i.e. after 2 transitions from AP17. This behavior is illustrated by the plot for $q^2_{17,1}(t)$ in Figure 4.1. Hence, for transitions that are destined to AP1, a user that initially has its active session associated with AP17 will likely remain in the same location for a relatively short period of time before making a transition to either AP1 directly, or to other locations before ending up

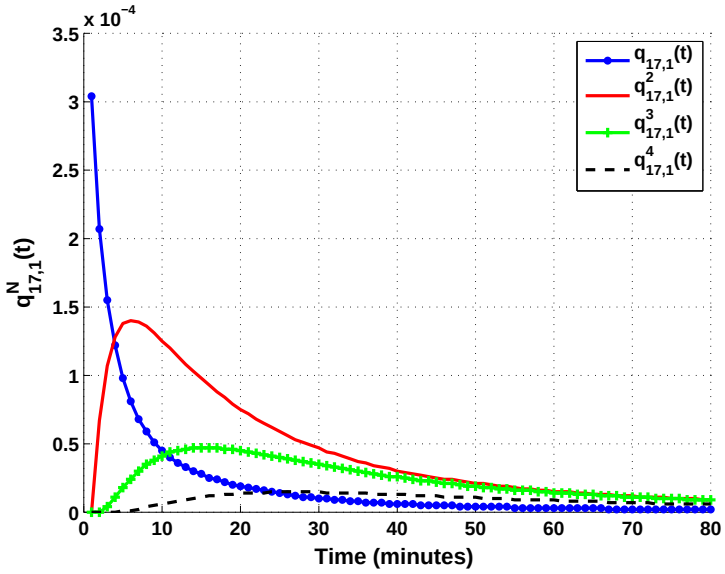


Fig. 4.1: A plot of the multi-transition prediction results $q^N_{17,1}(t)$, from AP17 to AP1.

in AP1.

A user making a transition from AP17 to AP1 after N transitions may also include the instances where the user re-visits AP1 after completing multiple transitions. For example, the probabilities $q^3_{17,1}(t)$ and $Q^3_{17,1}(t)$ also include the possibility of making the following sequence of transitions AP17 $\rightarrow$ AP1 $\rightarrow$ AP17 $\rightarrow$ AP1. This type of behavior is due to the user moving back and forth between the two locations. Thus, $q^3_{17,1}(t)$ should not be interpreted as the probability of having visited AP1 only *for the first time* and after 3 transitions from AP17. The probabilities $q_{i,j}(t)$ can assist with identifying the mobility behavior of a user between two particular locations. They can also help with understanding the route a user might take between the two locations as well as the number of transitions needed between them. For example, the overall higher probability $q^2_{17,1}(t)$ in Figures 4.1 and 4.2 seems to imply that the route from AP17 to AP1 is most likely via another single access point. These results could also be used with determining the number of transitions that are likely needed before a user ends up having its connection associated with

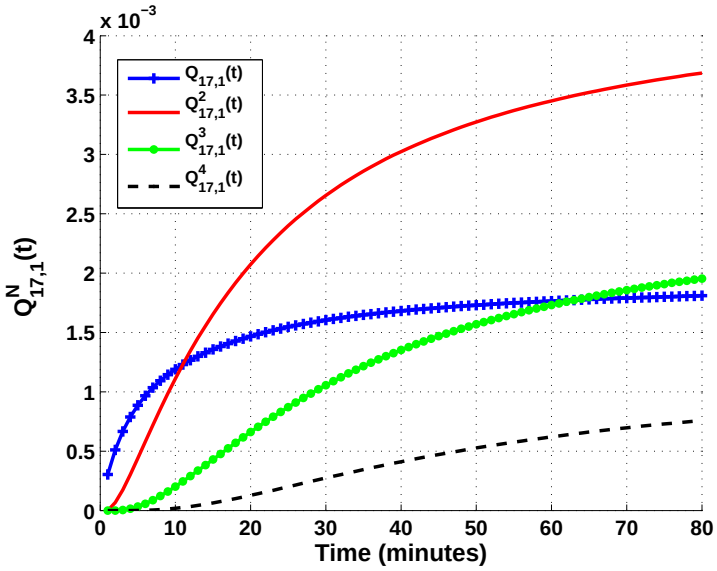


Fig. 4.2: A plot of the cumulative multi-transition prediction results $Q^N_{17,1}(t)$, from AP17 to AP1.

a particular access point. The process could ultimately be utilized for the purpose of making end-to-end connection predictions, along with assisting in the estimation of the future resources needed to sustain a user's ongoing active session during its multi-transition mobility.

Another example is reported in Figures 4.3 and 4.4. It looks at the likelihood of a mobile user making a transition from AP3 to its neighboring locations AP1, AP10, and AP21 on the same floor of the library building. The results show how the chances of ending up in a particular location can change with the number of transitions. For instance, Figure 4.3 shows how a user initially associated with AP3 is least likely to end up being associated with AP1 after 2 transitions. The conclusion changes if we were to estimate the same likelihoods after 3 transitions, as shown in Figure 4.4. This information can be crucial for the purpose of making end-to-end mobility predictions, in addition to understanding the likely number of transitions needed before ending up being associated with a particular access point.

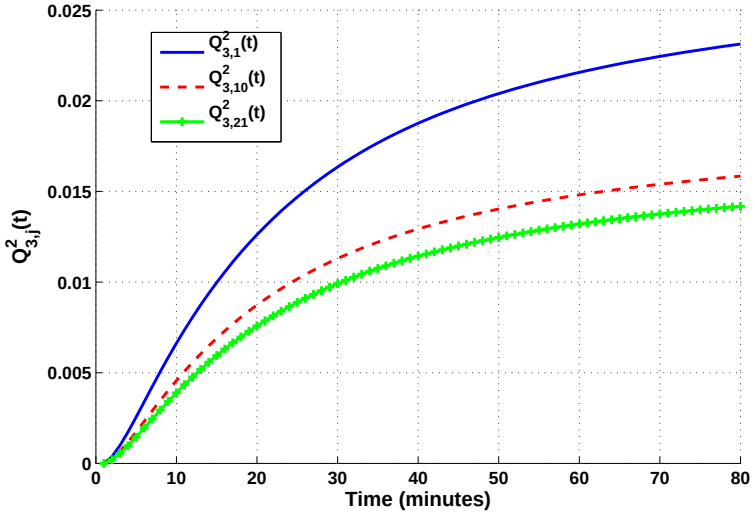


Fig. 4.3: A plot of the cumulative multi-transition prediction results $Q^2_{3,j}(t)$.

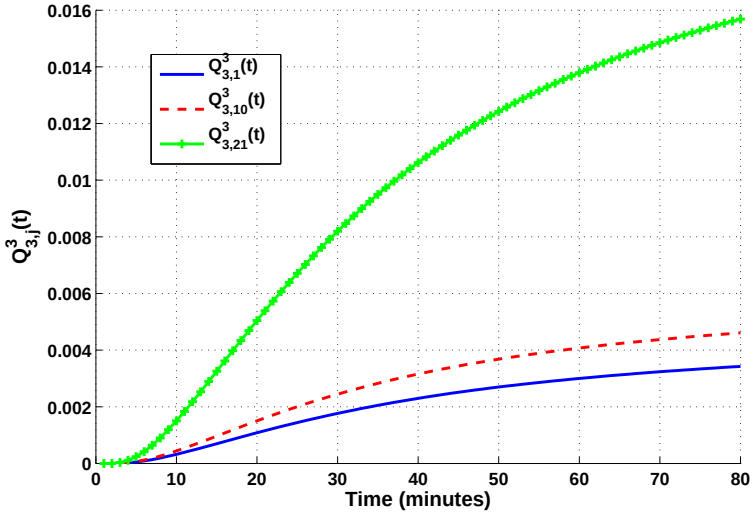


Fig. 4.4: A plot of the cumulative multi-transition prediction results $Q^3_{3,j}(t)$.

4.2 N-Transition Prediction Accuracy

For testing how well the proposed multi-transition prediction performs in comparison with some of the conventional methods, the same data set from [83] that was analyzed in Chapter 3 will be used again in this section. In order to directly apply Equation (4.3) for evaluating the accuracy of the multi-transition predictions, some of the parameters for the semi-Markov model were re-calculated, specifically the state sojourn time distributions. The element of the mobility model in Chapter 3 assumed a continuous-time distribution for the sojourn times, whereas the proposed computation for $\mathbf{q}^N(t)$ in Equation 4.3 assumes the state sojourn times $g_{i,j}(t)$ are discrete-time distributions. To facilitate this computation, it was chosen to fit these state sojourn times to a discrete-time phase-type distribution, i.e. $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$. The tool developed by Horváth and Telek in [74] was employed for this task. The distribution fitting process involved having to select the number of phases for estimating the parameters of the distribution. For simplicity, it was chosen to have each of the relevant state sojourn times fitted to a phase-type distribution with 4 phases. The sojourn times were also fitted to a geometric distribution, i.e. $G_{i,j}(t) \sim Geom(p)$, in order to examine whether the distribution of the sojourn times has a memoryless behavior and its impact on the prediction performance.

Similar to the approach taken in Chapter 3, the parameters for the proposed mobility model were processed using the first part of the data set that contained the events that had occurred prior to the year 2003. The mobility model was then used to process the remaining events and examine whether the predictions based on the results from $\mathbf{Q}^N(t)$ correspond to the actual events reported in the traffic data. For each event that has just entered state/location j after T minutes from entering its previous Nth state i, $\mathbf{Q}^N(t)$ was computed for $t \in [T - \Delta, T + \Delta]$. If the maxi-

mum probability returned by $\mathbf{Q}^N(t)$ corresponds with the event that was processed, i.e. $\max_k\{Q^N_{i,k}(t)\} = Q_{i,j}(t)$, then the prediction was said to be accurate. In this example, $\Delta = 1$ was chosen meaning that the events were checked against the computed likelihoods $\mathbf{Q}^N(t)$ within ± 1 of the actual temporal occurrence. The accuracy of the prediction results computed using Equation (4.4) were also compared with those conventional schemes that only employ the transition probabilities $\widetilde{P}_{i,j}$ and $P_{i,j}$ alone. The details for $\widetilde{P}_{i,j}$ have been previously covered in Equation (3.2). The Nth transition predictions using these transition probabilities alone were evaluated by simply computing $\widetilde{\mathbf{P}}^N$ and $\mathbf{P}^N$, respectively, where the matrices $\widetilde{\mathbf{P}} = \left\{\widetilde{P}_{i,j}\right\}$ and $\mathbf{P} = \{P_{i,j}\}$.

Tables 4.1 to 4.3 lists a summary of the average fraction of transitions that $\mathbf{Q}^N(t)$ was successful at predicting for each user and up to $N = 3$. These prediction accuracies were compared with those returned using $\mathbf{P}^N$ and $\widetilde{\mathbf{P}}^N$. The results for $N = 1$ were also included for comparison purposes and have been previously discussed in Section 3.2. They show a slight difference from those reported in Table 3.1 due to the different type of distributions used for the analysis in this section. The set of events that were parsed in the data set were limited to cover only those sessions that exhibited at least 2 location transitions before termination. This limitation had allowed for a more fair comparison between the results for different N.

For the case of $N = 1$, the results in the table agree with the conclusions made in Section 3.2. For $N = 2$, the $\mathbf{Q}^2(t)$ predictor has shown a better performance than the ones using the $\mathbf{P}^2$ and $\widetilde{\mathbf{P}}^2$ predictors in quite a few cases, e.g. transitions from AP14 and AP17. This is likely due to the static nature of the results returned by $\mathbf{P}^2$ and $\widetilde{\mathbf{P}}^2$ while the $\mathbf{Q}^2(t)$ predictor is influenced by the temporal behaviors. However, for the case of $N = 3$, all three predictors appear to have the same accuracy on the overall. This may be seen as the $\mathbf{Q}^3(t)$ predictor having the same performance

Tab. 4.1: Summary of N-transition prediction accuracy results, with $N = 1$, and for transitions from the access points in Library Building 2.

Transitions From AP	Conventional Predictors		Proposed Predictor Using MRP		Total # of Transitions	Total # of Users
	$P_{i,j}$ Predictor	$\widetilde{P}_{i,j}$ Predictor	$Q_{i,j}(t)$ Predictor $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$	$Q_{i,j}(t)$ Predictor $G_{i,j}(t) \sim Geom(p)$		
AP1	0.69918	0.69918	0.63587	0.47447	8541	2227
AP2	0.70159	0.70159	0.70159	0.47638	1342	321
AP3	0.21397	0.60156	0.34379	0.44917	14274	2550
AP4	0.63804	0.63804	0.63804	0.44993	8525	1469
AP5	0.75749	0.75749	0.75749	0.75749	5070	1140
AP6	0.69701	0.69701	0.66874	0.5401	8458	1713
AP7	0.7453	0.7453	0.7453	0.57036	16256	813
AP8	0.66841	0.66841	0.66841	0.53307	1500	485
AP9	0.84934	0.84934	0.84934	0.84934	16312	1915
AP10	0.10326	0.51308	0.17094	0.31562	9751	2158
AP11	0.85182	0.85182	0.85182	0.85182	18905	2363
AP12	0.61723	0.61723	0.61723	0.61723	2669	298
AP13	0.70512	0.70512	0.70512	0.62715	7271	1450
AP14	0.18531	0.61395	0.46797	0.47909	33936	2917
AP15	0.6069	0.6069	0.6069	0.6069	13003	185
AP16	0.084491	0.64882	0.13955	0.28496	9958	1260
AP17	0.20401	0.58419	0.31533	0.42923	26083	2818
AP18	0.078299	0.46463	0.12258	0.21103	981	118
AP19	0.67171	0.67171	0.67171	0.49827	19502	3066
AP20	0.45703	0.45703	0.51938	0.59495	2553	754
AP21	0.70726	0.70726	0.70726	0.51736	22708	3367
Overall	0.5354	0.6571	0.5669	0.5302		

Tab. 4.2: Summary of N-transition prediction accuracy results, with $N = 2$, and for transitions from the access points in Library Building 2.

Transitions From AP	Conventional Predictors $\mathbf{P}^2$ Predictor	Conventional Predictors $\widetilde{\mathbf{P}}^2$ Predictor	Proposed Predictor Using MRP $\mathbf{Q}^2(t)$ Predictor $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$	Proposed Predictor Using MRP $\mathbf{Q}^2(t)$ Predictor $G_{i,j}(t) \sim Geom(p)$	Total # of Transitions	Total # of Users
AP1	0.42935	0.42935	0.42935	0.37619	2406	944
AP2	0.65792	0.65792	0.65792	0.5041	263	118
AP3	0.45079	0.45079	0.45079	0.40697	6115	1345
AP4	0.27395	0.27395	0.27395	0.18239	2907	709
AP5	0.35514	0.35514	0.35514	0.35514	1492	368
AP6	0.34759	0.34759	0.34759	0.26574	2592	744
AP7	0.30289	0.30289	0.34828	0.37249	6939	307
AP8	0.36925	0.36925	0.36925	0.36925	595	184
AP9	0.28409	0.28409	0.41698	0.4373	2866	607
AP10	0.34462	0.34462	0.34462	0.28934	4978	1282
AP11	0.34976	0.34976	0.34976	0.37577	2584	790
AP12	0.46576	0.46576	0.46576	0.46576	478	142
AP13	0.47131	0.47131	0.47131	0.47131	1849	598
AP14	0.35662	0.35662	0.50308	0.50596	14575	1705
AP15	0.41902	0.41902	0.41902	0.28355	4969	120
AP16	0.36143	0.36143	0.36143	0.36517	5026	512
AP17	0.35833	0.35833	0.49338	0.5145	13117	1718
AP18	0.52128	0.52128	0.26633	0.33952	337	69
AP19	0.47995	0.47995	0.46012	0.24486	6429	1578
AP20	0.59603	0.59603	0.59603	0.6048	1508	462
AP21	0.42011	0.42011	0.44179	0.42186	6707	1623
Overall	0.4102	0.4102	0.42009	0.3882		

Tab. 4.3: Summary of N-transition prediction accuracy results, with $N = 3$, and for transitions from the access points in Library Building 2.

Transitions From AP	Conventional Predictors $\mathbf{P}^3$ Predictor	Conventional Predictors $\widetilde{\mathbf{P}}^3$ Predictor	Proposed Predictor Using MRP $\mathbf{Q}^3(t)$ Predictor $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$	Proposed Predictor Using MRP $\mathbf{Q}^3(t)$ Predictor $G_{i,j}(t) \sim Geom(p)$	Total # of Transitions	Total # of Users
AP1	0.32787	0.32787	0.32787	0.30844	1415	538
AP2	0.39943	0.39943	0.39943	0.25874	125	48
AP3	0.29373	0.29373	0.29373	0.28736	4032	835
AP4	0.28952	0.28952	0.28952	0.20919	2206	525
AP5	0.31739	0.31739	0.31739	0.31739	929	222
AP6	0.32515	0.32515	0.32515	0.37725	1788	516
AP7	0.36594	0.36594	0.36594	0.38054	4847	216
AP8	0.30318	0.30318	0.30318	0.21672	284	110
AP9	0.32287	0.32287	0.32287	0.41365	2237	417
AP10	0.26288	0.26288	0.26288	0.25489	3714	874
AP11	0.32378	0.32378	0.32378	0.32378	1704	519
AP12	0.27784	0.27784	0.27784	0.27784	256	89
AP13	0.30194	0.30194	0.30194	0.30475	1148	356
AP14	0.29984	0.29984	0.29984	0.37974	11079	1252
AP15	0.32321	0.32321	0.32321	0.33297	3203	87
AP16	0.26556	0.26556	0.26556	0.20769	3259	291
AP17	0.3213	0.3213	0.3213	0.39546	10058	1224
AP18	0.18163	0.18163	0.18163	0.17032	160	38
AP19	0.36752	0.36752	0.36752	0.31091	4137	956
AP20	0.20576	0.20576	0.5571	0.54625	789	212
AP21	0.28672	0.28672	0.28672	0.24998	4394	1064
Overall	0.303	0.303	0.3197	0.3107		

as those of the $\mathbf{P}^3$ and $\widetilde{\mathbf{P}}^3$ predictors which is not entirely accurate. The $\mathbf{Q}^3(t)$ predictor also included the time window at which the event occurred and this extra information was not returned by the other two predictors. These results could lead to the conclusion that the performance of the semi-Markov predictor diminishes with increasing N and this appears to be true for the data set [83] that was analyzed in this section. The results could also suggest that the temporal information may not have any significant influence in the case of generating predictions with higher N.

Figures 4.5 and 4.6 highlights the difference in the accuracies achieved by using the $\mathbf{Q}^N(t)$ predictor, with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and $G_{i,j}(t) \sim Geom(p)$, and the $\mathbf{P}^N$ predictor, for $N = 2$ and $N = 3$, respectively. In Figure 4.5, the accuracy of the $\mathbf{Q}^2(t)$ predictor with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ were in many cases higher than those achieved with $G_{i,j}(t) \sim Geom(p)$. This shows that the use of the distribution with the non-memoryless behavior had helped with achieving a higher prediction accuracy for a lot of the transitions and for the case of $N = 2$. The same can be said for the case of $N = 3$, as shown by Figure 4.6. Since there were a few instances where the use of the geometric distribution offered a much higher prediction accuracy, the network managers do not have to constrict themselves to using a single type of distribution for all their predictions. For example, they could select the $\mathbf{Q}^N(t)$ predictor with $G_{i,j}(t) \sim Geom(p)$ for transitions from AP9 due to the higher accuracy shown in the results, while the $\mathbf{Q}^N(t)$ predictor with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ may be more suitable for transitions from AP10. This illustrates one of the further advantages of applying the proposed prediction scheme.

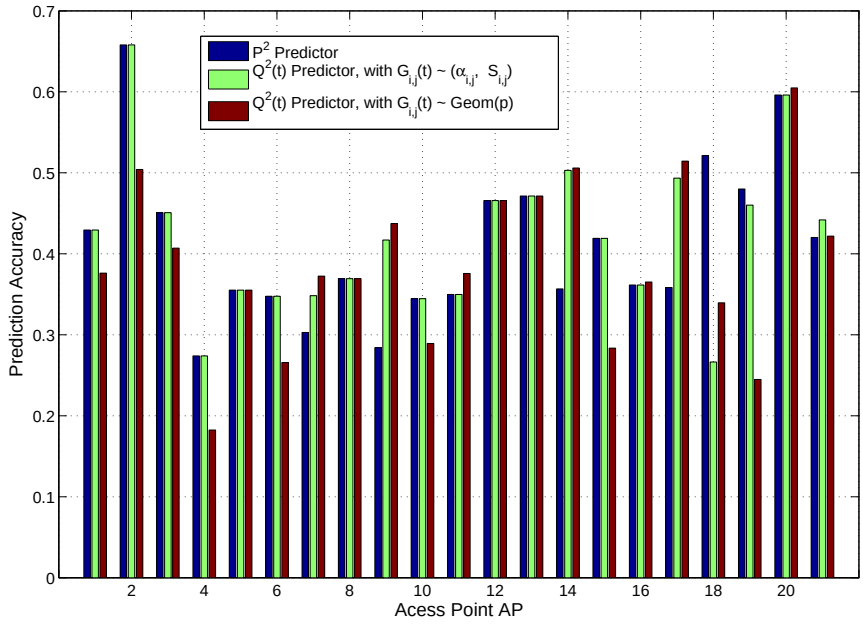


Fig. 4.5: The performance of the $\mathbf{Q}^2(t)$ predictor with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and $G_{i,j}(t) \sim Geom(p)$, and compared with the $\mathbf{P}^2$ predictor.

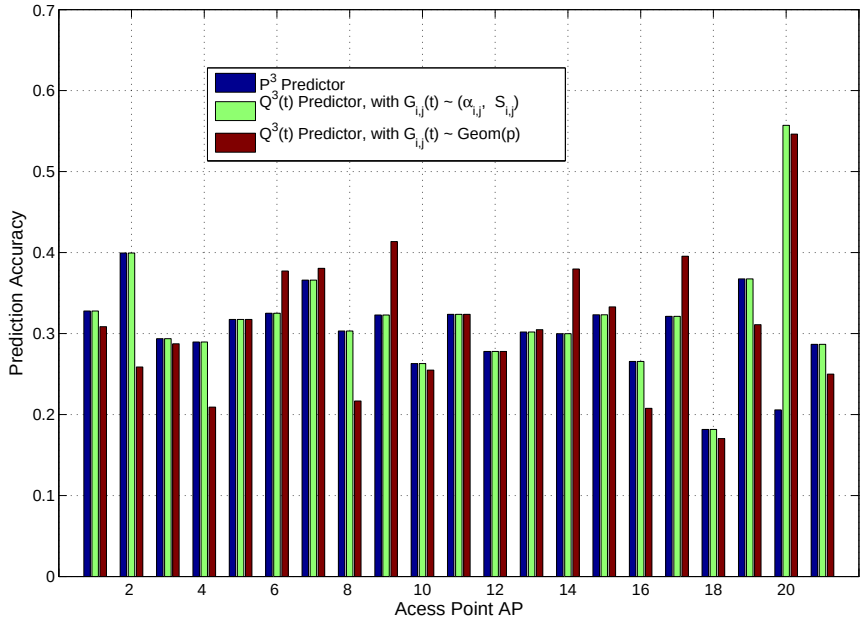


Fig. 4.6: The performance of the $\mathbf{Q}^3(t)$ predictor with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and $G_{i,j}(t) \sim Geom(p)$, and compared with the $\mathbf{P}^3$ predictor.

4.3 Spatial-Temporal Traffic Estimation

An understanding of how the number of users in a network varies with time can further provide network managers with an insight of the traffic demands at each location in the network. Such information can assist with optimizing the allocation of the network's resources. In this section, we will only focus on the case where an equal amount of resources are required for each of the users in the network, with the difference being in the amount of time spent utilizing the resources. An example of such a network are those that support voice calls only, whereby each connection is given a distinct channel that is equal in bandwidth with all the other available channels. Hence, the network traffic load can be studied by estimating the number of users that have active sessions at each location. The traffic of users with idle sessions can also be investigated since they may *potentially* acquire access to the network.

The proposed mobility model can be employed to further predict the number of users at each of the locations by estimating the average number of users that are expected to transition from one location to another and within a time period t. Let V_{-i} and V_i denote the current (i.e. at $t = 0$) number of users with idle and active sessions in location i of the network, respectively. Let us further define $Y_i(t)$ as the expected number of active users transferring into location i within time t. This consists of the proportion of users, with either active or idle sessions, that are expected to transfer from all locations j into i within time t and can end up in state i after N transitions. This also includes the proportion of users that are expected to remain in the same state i within time t. Note that these N transitions could include switching between being idle and active while remaining in the same location, as well as entering and returning to state i more than once. Hence, the

result for $Y_i(t)$ can be evaluated as follows

$$Y_i(t) \;=\; \sum_j V_j \sum_{n=1}^{t} Q_{j,i}^n(t), \tag{4.6}$$

with states j being all those locations that lead to state i after 1 or more transitions, and for users with both idle and active sessions. They include those that transition from state i and return to the same state after $N \geq 2$ transitions. A similar definition can be given for $Y_{-i}(t)$ for the case of idle users. Similarly, we can also compute

$$y_i(t) \;=\; \sum_j V_j \sum_{n=1}^{t} q_{j,i}^n(t), \tag{4.7}$$

which describes the number of users that are expected to transfer into state i *at* time t.

If we define the vector $\mathbf{V}$ such that $\mathbf{V} = \{(V_{-1}, V_{-2}, \cdots, V_{-L_I}), (V_1, V_2, \cdots, V_{L_A})\}$, then we can write the following,

$$\mathbf{Y}(t) \;=\; \mathbf{V} \sum_{n=1}^{t} \mathbf{Q}^n(t) \;=\; \mathbf{V}\mathbf{\Lambda}^t(t) \tag{4.8}$$

where $\mathbf{Y}(t)$ is a vector of the expected number of both idle and active users that are predicted to be transferred into the various locations within time t. A similar definition could also be given for the case of estimating the expected number of idle and active users that are expected to be transferring into the various locations at time t, i.e. computing $\mathbf{y}(t)$ from $\mathbf{q}^N(t)$. Computing $\mathbf{Y}(t)$ can be quite cumbersome, especially when the dimensions of $\mathbf{Q}(t)$ are large. Nevertheless, the results from $\mathbf{Y}(t)$ can be altogether beneficial, as demonstrated in the examples given next.

To illustrate the benefits of such computations, we will again continue to utilize the same subset of the traffic traces used in the previous sections. However, the data

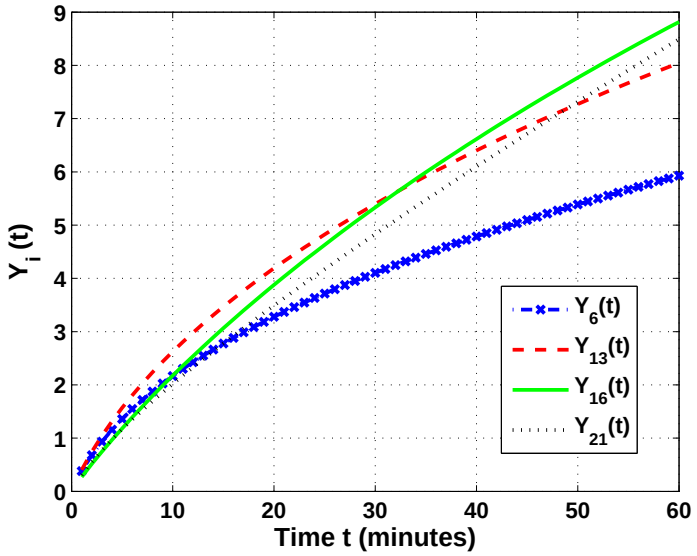


Fig. 4.7: The expected number of users transferred into locations served by AP6, AP13, AP16, and AP21, within t minutes.

had not kept track of the users in the various locations with idle sessions. Hence, and for simplicity, we will assume the example where the users in the network are always active and the transitions between the access points only involve those users with active sessions. Moreover, in this example the initial number of users $\mathbf{V}$ associated with each of the 21 different access points were chosen randomly and for illustrative purposes.

Figure 4.7 shows on average the number of users that are expected to transition and associate themselves with the access points AP6, AP13, AP16, and AP21, within t minutes from the current population distribution $\mathbf{V}$. Each of these access points are known to be on separate floors of the library building. The results show that within the first 30 minutes from the current time at which the population of users is $\mathbf{V}$, more users on average are expected to transition and associate their connections with AP13 when compared with the remaining 3 access points. Note that this can be due to a single or multiple transitions by each user before finally associating their

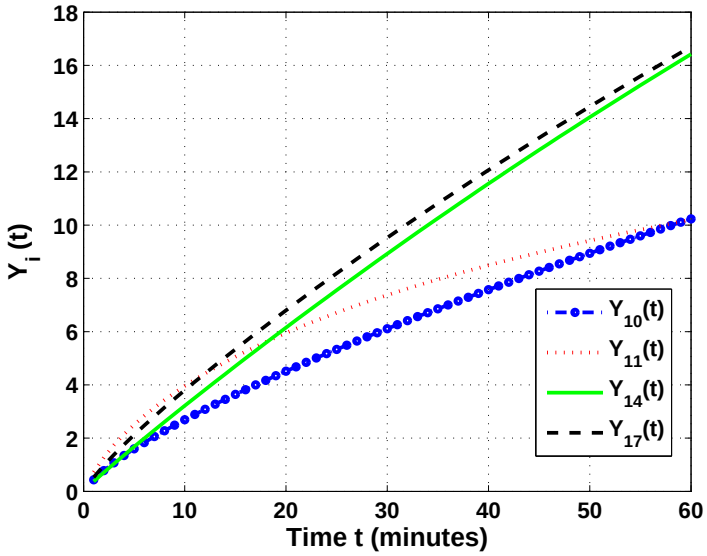


Fig. 4.8: The expected number of users transferred into locations served by AP10, AP11, AP14, and AP17, within t minutes.

connections with AP13. Furthermore, $Y_{13}(t)$ is the expected number of users that have transferred into state 13 within t minutes. This is not to be interpreted as the expected number of users that are to be found in state 13 within t minutes since some of them may have transitioned out of state 13 by t minutes. The higher $Y_{13}(30)$ when compared with the other results suggest that the location served by AP13 attracts more users than the others during that particular time period. However, for larger time periods, AP16 seems to be the one with the higher number of expected users transitioning to it. Hence, AP13 attracts more users on the short-term whereas AP16 attracts more users on the long-term. This kind of information can be quite valuable to network managers for the purpose of provisioning sufficient network resources for its users.

Another example is given in Figure 4.8 which shows a similar behavior among a different set of access points. In this example, AP14 and AP17 are located on the same floor and are expected to have more users transitioning towards them

when compared with the other two access points that are each located on adjacent floors. The results for $Y_{11}(t)$ indicate that AP11 is expected to have more users transitioning towards it within the first 10 minutes when compared with the other access points. However, AP11 is seen to have much fewer users transitioning to it when making predictions further ahead in time. A network manager could assume from these predictions that less resources are needed for AP11 after the first 10 minutes and could instead direct the excess resources from AP11 to AP14 and/or AP17.

The computations from Equation 4.8 assume that the network can serve an unspecified number of users with active sessions at each location. In reality, this is not true since the number of active connections at any given time period is limited by the capacity of the network at each of the locations. This is due to the network's limited resources. Hence, the model can be further modified to consider a given maximum number of active connections that can be served by the network at each location. The maximum number of active connections is limited, whereas the capacity of users with idle sessions can be assumed to be as large as the total population of users in the network. Certain applications may also require having to impose a limit on the number of users with idle sessions at a given location. Such cases will not be considered in this thesis.

Let C_i be the maximum number of active sessions that can be allowed by the network to be transferred at each location i. Therefore, active sessions that are transferred into a location i which is serving at full capacity are prematurely terminated by the network and included in the number of users with idle connections in the same location. This is equivalent to having a user's active session being dropped after attempting a handoff to the new location. Define $\widehat{Y}_i(t)$ and $\widehat{Y}_{-i}(t)$ as the expected number of users that are expected to have their active and idle

sessions transferred within time t into location i with a limited serving capacity of C_i, respectively. Using Equation (4.8), $\widehat{\mathbf{Y}}(t)$ can be computed from $\mathbf{Y}(t)$ as follows,

$$\widehat{\mathbf{Y}}(t) \;=\; \left[\widehat{Y}_i(t)\;,\; \widehat{Y}_{-i}(t)\;:\;\forall i\right], \tag{4.9}$$

$$\text{where} \quad \widehat{Y}_i(t) = \min\left\{Y_i(t), C_i\right\},$$

$$\text{and} \quad \widehat{Y}_{-i}(t) = \max\left\{Y_{-i}(t)\;,\; Y_{-i}(t) + Y_i(t) - C_i\right\}.$$

Not only can such results predict the spatial and temporal traffic load at each location, but they could also be used to estimate how far ahead in time the network is likely to drop any further incoming sessions. Such information could be used by the network manager to administer some alternative call admission control.

4.4 Network Resource Reservation

One of the major applications of mobility prediction is to estimate the amount of resources that need to be reserved at the neighboring locations in the network. This allows for a user's ongoing and active session to be uninterrupted while being mobile between the various locations. In this section, we show how the proposed mobility prediction can generally be applied for the purpose of advanced resource reservation in wireless networks. The time window in which these reservations need to be maintained are also considered. Examples of how such prediction results can be applied for resource reservation purposes can be seen in the works done by Choi and Shin in [22] and Kim in [86].

Define σ as the amount of network resources that are required for each mobile connection/user, e.g. $\sigma = 1$ channel for each connection in a voice cellular network. It is assumed that σ is the same for all users. Thus, the expected amount of network resources that need to be reserved for future demands at location i and within time t is given as $R_i(t)$, such that

$$R_i(t) \;=\; \sigma \sum_j V_j \sum_{n=1}^{t} Q_{j,i}^n(t) \;=\; \sigma Y_i(t), \tag{4.10}$$

where V_j is the number of active sessions in location j at time $t = 0$. Equation (4.10) considers both the active and idle users in all locations that are predicted to transition into location i in N or fewer transitions (with $N \leq t$). Hence, the network should, on average, be expected to reserve up to $R_i(t)$ resources at location i within time t. Alternatively, we can define the following,

$$r_i(t) \;=\; \sigma \sum_j V_j \sum_{n=1}^{t} q_{j,i}^n(t) \;=\; \sigma y_i(t), \tag{4.11}$$

where $r_i(t)$ is the expected amount of network resources that are needed for users arriving into location i and *at* time t. Note that the results for $R_i(t)$ and $r_i(t)$ simply rely on the computations of $Y_i(t)$ and $y_i(t)$ from the previous section. This form of resource reservation can be applicable in situations where the network manager may choose to reserve the aggregate expected amount of resources needed for a group of users that may arrive at a particular location within a certain time period. The allocation of the resources may instead be done for each individual user on a per-transition basis and the reservation can be based simply on the predictions returned by the $Q_{i,j}(t)$ predictor.

The results for $R_i(t)$ (and $r_i(t)$) account for users that have both idle as well as active and ongoing sessions transferred from one location to another. Note how these results include those users that are initially idle in terms of network usage. The reason is that such users can potentially be active in the given location and influence the resources that need to be reserved. Alternatively, if one wished to only consider the active users, then the $Q_{i,j}(t)$ elements in the model need only be defined for the case of transitions with active sessions alone.

The choice of reserving $R_i(t)$ amount of resources is an example of a fractional-based resource reservation scheme. In other words, the result of $R_i(t)$ can assume any real number. For example, the network manager could be recommended to reserve 3.37 units of resources by the predictor. In such a case, the network manager may choose to round up/down to the nearest integer number of resources to be reserved. This is done especially when reserving a fraction of a resource unit is not possible.

4.5 Simulation of Network Resource Reservation

Many of the proposed resource allocation methods rely on steady/invariant user-behavioral statistics for determining both the location and the amount of network resources to be reserved. This type of reservation is specifically for those users that may potentially handover their ongoing active sessions at the neighboring access points. Such reservations tend to reduce the chances of a mobile user's ongoing session being prematurely terminated by the network due to insufficient resources at the newly associated access point. A wireless network which incorporates a resource reservation scheme into its operations has a strong interest at minimizing the connection dropping rates and thereby improving the overall customer satisfaction. However, reserving a pool of resources for future handover connections usually comes at the risk of having less resources available for new connections. This would potentially increase the new connection blocking rates. Hence, a suitable resource reservation scheme would tend to reach a compromise between the blocking and dropping rates.

A virtual wireless networking environment was developed to simulate the mobility and activity of a number of independent users within a certain coverage area. This was done to assess whether the proposed mobility prediction can perform better than the conventional schemes (using the transition probabilities alone), in terms of reserving the necessary resources for handover connections. The simulation was initially run to first collect the necessary data to acquire the statistics needed to construct the model for the mobility prediction. The data include the location transition probabilities and the location sojourn times. The next step involved applying the mobility prediction for resource reservation purposes and monitoring the network's performance in terms of the connection blocking and dropping rates.

The majority of researchers constructed their mobile network simulations based on the various common methods and techniques summarized by Camp et al. in [90]. However, many of these models suffer from quite unrealistic assumptions such as sudden and drastic changes in speed and/or direction of mobility per wireless node. Some of these problems have been identified by Yoon et al. in [91] and Theoleyre et al. in [92]. Bettstetter [93] proposed a stochastic and "smooth" mobility model to improve on the previously unrealistic assumptions. His mobility model introduces a gradual change in the speed and direction of a wireless node which depends on the node's acceleration, deceleration and turning speeds. This type of stochastic mobility model was further extended by Zhao and Wang in [94] to one that is based on a semi-Markov process for a more general mobility behavior. Other mobility models that follow a similar idea to the one given by Bettstetter in [93] are those proposed by Yoon et al. in [95], Boudec and Vojnovic in [96], and Hsu et al. in [97].

To simplify matters, we developed a discrete-event simulation in Matlab [82] and adopted the "Smooth Mobility Model" proposed by Bettstetter in [93] for characterizing the mobility behavior of each user in the network. The main idea behind this mobility model is that both the changes in speed and direction are controlled by random processes which are dependent on each other. Each node's mobility path is governed by a series of assigned target speeds and directions. At any time t, a node with a given target speed of v_t and target direction d_t (in radians) accelerates/decelerates towards the chosen target. The node remains on the same course until either a new target $v_{t+\tau_v}$ or $d_{t+\tau_d}$ is selected at times $t+\tau_v$ and $t+\tau_d$, respectively. The times between the re-assignment of the new target speed and direction, i.e. τ_v and τ_d, are assumed to be random and follow a Poisson process. The new target speeds and directions are uniformly selected and the mobile nodes in the network are assumed to be homogeneous. The random times for each node's active and

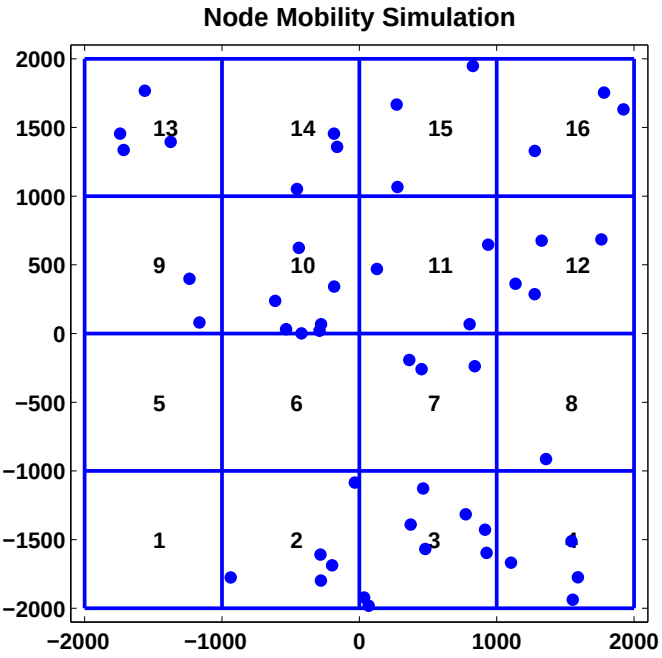


Fig. 4.9: Simulation environment for network resource reservation.

idle session times are assumed to be geometrically distributed. A 16−cell network was constructed with each cell being administered by a single access point that can only handle a limited number of active sessions at any given time. A wrap-around structure [98] for the 16−cell network was implemented to avoid using a bounded simulation area that tends to bias the output statistics and renders them inaccurate.

Figure 4.9 shows a snapshot of the 16−cell network that was implemented in the simulation, with the points on the grid representing the positions of the wireless nodes at the current time. A node that exits to the right of cell 4 will transition towards the left side of cell 1 and those leaving from the bottom of cell 4 will transition into cell 16 from the top side. This is how the boundary effect is eliminated and the same is true for the nodes departing from the other boundary cells. The parameters that were used in the simulation are given in Table 4.4. Some were chosen based on the numerical examples given by Bettstetter in [93].

Tab. 4.4: Summary of parameters for the simulation of network resource reservation.

Parameter	**Value(s)**
Mean Time for Speed Change	$50s$
Mean Time for Direction Change	$85s$
Choice of Speeds	$\{0, 4, 8, 10, 14\}$ m/s
Acceleration	2.5 m/s^2
Deceleration	4 m/s^2
Choice of Direction	$[-\pi, \pi]$
Mean Session Idle Time	$500s$
Mean Session Active Time	$100s$
Maximum Number of Active Sessions per Cell	4
Maximum Number of Reserved Channels per Cell	2

The simulation was initiated by first assigning 100 nodes randomly and uniformly across the 16 cells and allowing them to roam across the different locations for 5 million seconds of simulated run-time. During this time, the users' sessions are switching between being idle and active. In the first stage of the simulation, the necessary statistics were collected to compute the required parameters for $Q_{i,j}(t)$, for all (i, j) pairs. These parameters were only accumulated for those transitions that involved nodes with active sessions. The next stage of the simulation involved running it while executing resource reservation for handover connections using only $P_{i,j}$ at the first instance, and then repeating the same run but using $Q_{i,j}(t)$ instead. The former case is the conventional method while the latter considers the proposed MRP-based mobility model. This stage was run for another 5 million seconds of simulated run-time while computing the blocking and dropping rates at each cell.

For creating the resource reservations, each node was assumed to require a single unit of resource to continue its ongoing active session at the neighboring cell. The overall number of resources allowed to be reserved at each cell is restricted to a certain limit, which in this example is 2 channels (see Table 4.4). For the case

where only $P_{i,j}$ is used, a prediction is executed the moment a node either initiates a new session or transfers its active session into a given cell. A resource is reserved for that node at the next most likely location based on the generated prediction. If the node happens to cross into the cell with the reserved resource, the handover is completed and the resource prediction process is repeated. Otherwise, the reservation is freed at the moment the node has made a transition into a new location and the node with the handoff request will only be completed if a non-reserved resource is available for it. A similar approach was taken for the other case which applied the model with $Q_{i,j}(t)$ except that the probabilities are evaluated every Δ seconds from the moment a node enters a given location, and the resource reservation is executed accordingly. In the simulation, $\Delta = 10$ seconds was assumed such that if no handover was executed at the nth instant then a new prediction was evaluated, i.e. $[Q_{i,j}((n+1)\Delta) - Q_{i,j}(n\Delta)]$, and the resource reservation was modified. The re-evaluation of the prediction continues every Δ seconds until the node departs the current cell or completes its session. Other methods such as fractional-based resource reservation schemes discussed earlier in this section could also have been applied in the simulation. However, this basic approach was chosen to be applied for its simplicity.

Figures 4.10 and 4.11 compare the blocking and dropping probabilities using both types of predictions for reserving resources at each of the cells. On the whole, the results show the proposed predictor using $\mathbf{Q}(t)$ performed much better than the conventional predictor using $\mathbf{P}$, as summarized in Table 4.5. The predictor using the proposed MRP with $\mathbf{Q}(t)$ reduced both the overall blocking and dropping probabilities by 36.7% and 35.3%, respectively. In a few locations, the $\mathbf{P}$-based predictor performed better than the other method, e.g. cell 10. A closer look at the results revealed that these cells are almost always heavily loaded with traffic

which can be an undesirable situation for adaptive resource reservation schemes, as observed by Kim in [86]. Nevertheless, the results favor the $\mathbf{Q}(t)$-based resource reservation prediction overall in terms of both the blocking and dropping rates. The improvement could be attributed to the fact that the predictions using $\mathbf{Q}(t)$ advised the reservation of resources in the neighboring cells for shorter periods of time. This tended to free up more of the resources for other uses at any given time.

Tab. 4.5: Summary of results from simulation of network resource reservation.

	Using $\mathbf{P}$	Using $\mathbf{Q}(t)$	% Change
Overall Blocking Probability	0.141	0.0892	−36.7%
Overall Dropping Probability	0.0949	0.0613	−35.5%
Overall Prediction Accuracy	0.2754	0.5234	+90.1%

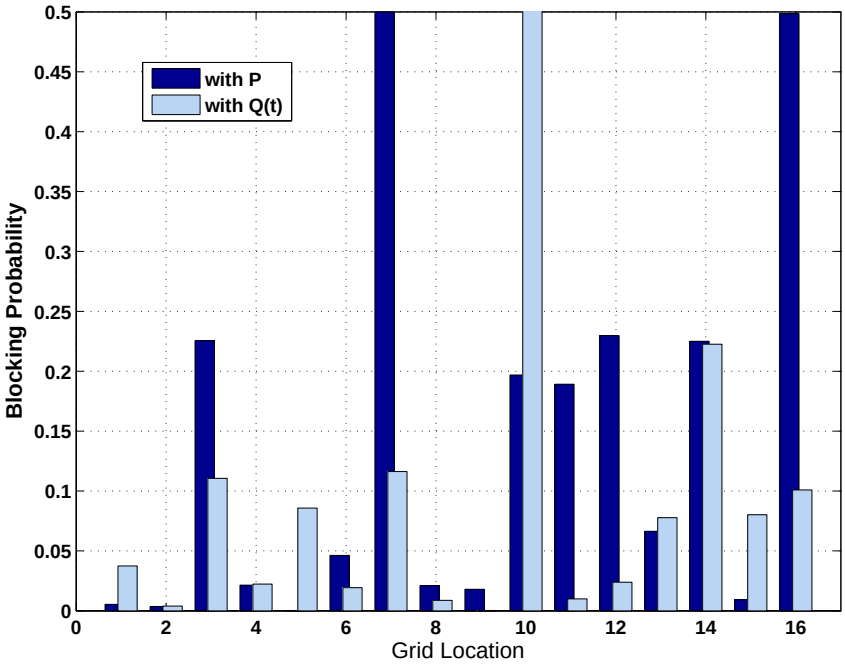


Fig. 4.10: Blocking probabilities per location, from the simulation for network resource reservation.

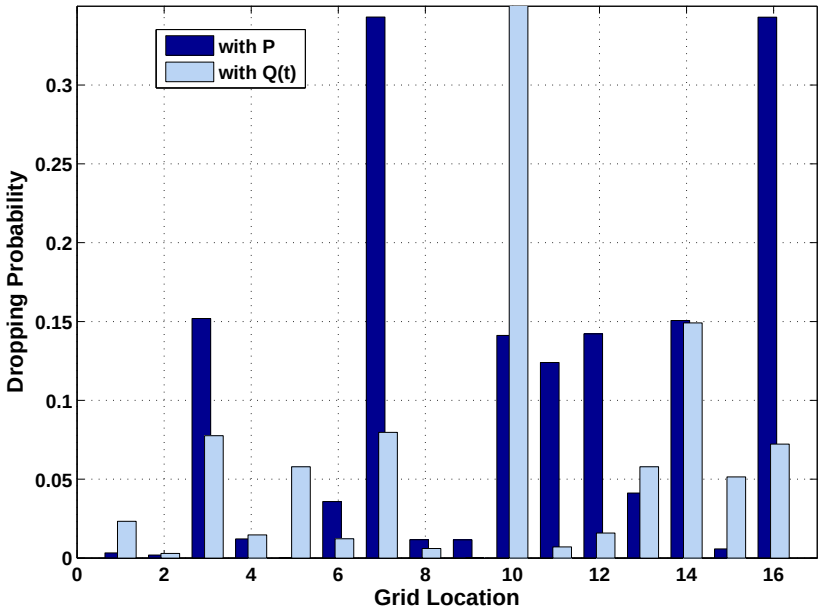


Fig. 4.11: Dropping probabilities per location, from the simulation for network resource reservation.

4.6 End-Location Predictions

Many researchers have identified the difficulty associated with predicting when and where a mobile user will end its network session. A user's session may undergo a random number of location transitions, including the possibility of the session terminating in the same location where it was initiated. It is certainly much easier to predict a user's subsequent transition than the state at which the session is terminated. In knowing when and where a session will complete, given the initial session-initiation state, the network manager could interpolate the surrounding locations that a mobile user may visit while undergoing an active session from start-to-end. In this case, end-location predictions apply for mobile users with sessions that remain active until they terminate. These predictions are particularly useful for scheduling the necessary resources that are required to be reserved at the various access points. Such reservations can assist with sustaining a mobile user's ongoing active session from start-to-finish and without any inconvenient disruptions. Note that this type of prediction is not the same as end-to-end predictions (see [87] and [68]). The latter is generally involved with determining the path details from one end to the other, whereas the end-location predictions are concerned with estimating when and where a session is likely to terminate. Both types of predictions depend on where the session was initiated.

For generating the end-location predictions, we apply the idea of absorbing Markov chains (see [88]) to our MRP-based mobility model. Consider the matrix $\hat{\mathbf{q}}(t)$ with the following structure,

$$\hat{\mathbf{q}}(t) \;=\; \begin{pmatrix} \mathbf{q_a}(t) & \mathbf{q_0}(t) \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{4.12}$$

$\mathbf{q_a}(t)$ is an $L_A \times L_A$ square matrix with elements $0 \leq q_{i,j}(t) \leq 1 \ \forall t$ and $i \neq j$, and represents the transitions from location i to j with an active session. $\mathbf{q_0}(t)$ is also an $L_A \times L_A$ square matrix with the elements $q_{i,-i}(t)$ on the diagonal that represent the transitions involving the termination of a session in location i at time t. $\mathbf{I}$ is an identity matrix.

Define $\mathbf{q_e}(t)$ as a row vector with the elements being the joint probabilities in which a mobile user terminates its ongoing active session in a particular location t units of time after entering the current location. These probabilities are conditioned on the current location of the mobile user. Using the definition for $\hat{\mathbf{q}}(t)$, and assuming that the state sojourn time distributions are of discrete form, we can compute $\mathbf{q_e}(t)$ as follows,

$$\mathbf{q_e}(t) \ = \ \alpha \mathbf{q_0}(t) \ + \ \alpha \sum_{\tau=1}^{t-1} \sum_{n=1}^{t-\tau} \mathbf{q_a}^n(t-\tau)\mathbf{q_0}(\tau). \tag{4.13}$$

The vector $\alpha = e_i'$ is a row of all zeros with a single entry of a 1 in location i and represents the state of the current location of a user with an active session. The matrix $\mathbf{q_a}^n(t-\tau)$ can be computed from $\mathbf{q_a}(t)$ and using Equation (4.3). A mobile user may terminate its current session with the network without moving out of the initial location, as given by probabilities in $\alpha\mathbf{q_0}(t)$. Otherwise, the user could move away from the initial location and terminate its session elsewhere, which occurs with a probability given by the second part of the expression for $\mathbf{q_e}(t)$. The formula given in Equation (4.13) is valid for the case where the state sojourn times assume a discrete distribution. A similar approach can be taken for developing this same metric for the case where the state sojourn times assume a continuous distribution.

Using the same data [83] for the network examined in Chapter 3, the results of $\mathbf{q_e}(t)$ for those users that have initiated their sessions at AP14 and AP17 in

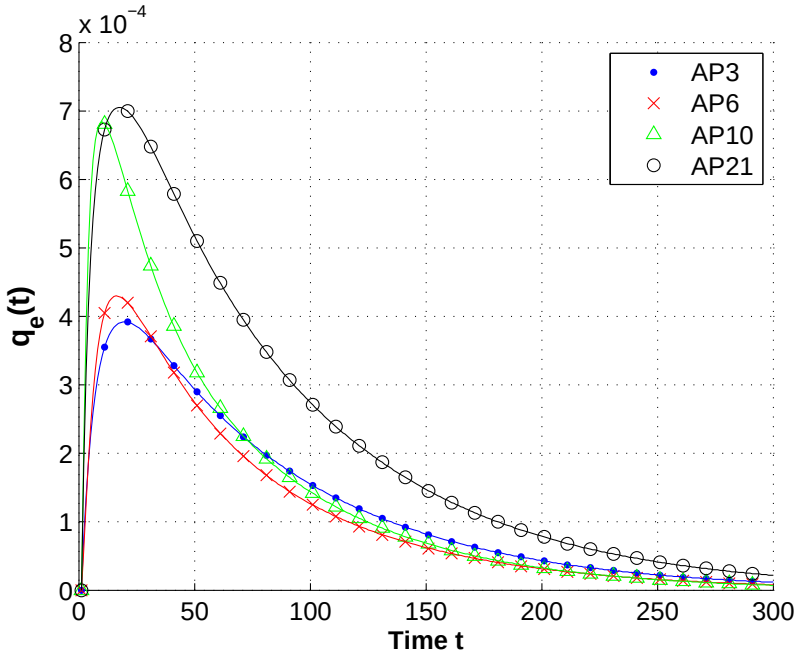


Fig. 4.12: A plot of the End-Location $\mathbf{q_e}(t)$ probabilities for mobile users having initiated their sessions at AP14.

Library Building 2 will be given next. The state sojourn times will again be fitted to a discrete-time phase-type distribution using the tool developed by Horváth and Telek in [74]. This in turn allowed for computing $\mathbf{q_e}(t)$ using Equation (4.13).

Figures 4.12 and 4.13 show the probabilities of a mobile user terminating its ongoing active session at the various locations within the same building, given that the session was initiated at AP14 and AP17, respectively. From the data set in [83], there are 21 distinct access points that cover the various locations in the building. Hence, there are 21 different locations in which a user can terminate its session within the same building. There were also many situations where a user may transition outside the building while continuing its active session. For simplicity, it was again chosen to limit the numerical analysis for transitions that occurred within the same building. Even though there should be 21 distinct plots in each of Figures 4.12 and 4.13, only the top 4 possibilities were shown in each for clarity.

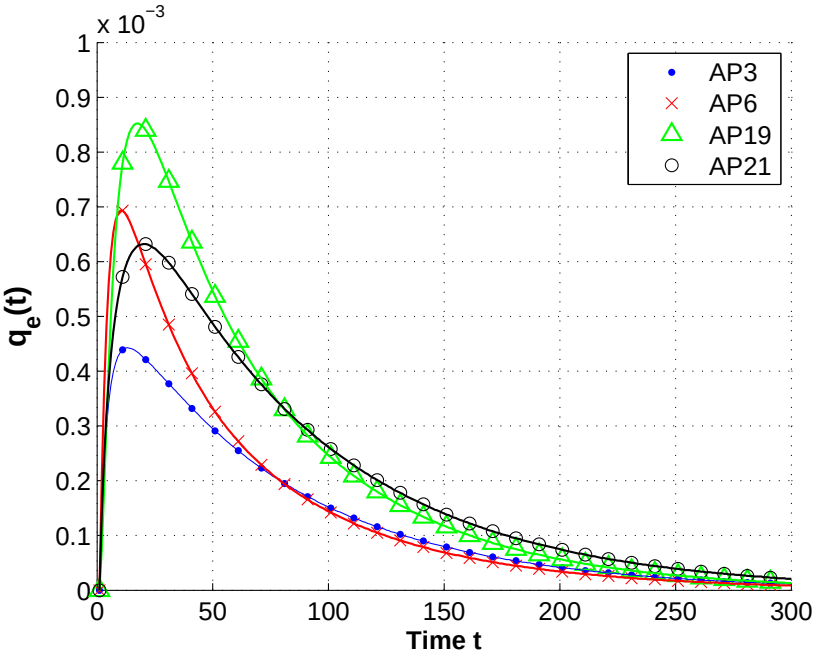


Fig. 4.13: A plot of the End-Location $\mathbf{q_e}(t)$ probabilities for mobile users having initiated their sessions at AP17.

Figure 4.12 shows how the probability of ending up in either of the 4 access points varies with the session time t and for the case where the session was initiated at AP14. During the duration t, the mobile user may transition into various locations along the way before terminating its session at a particular location. The results show that out of the 4 possibilities, a mobile user is more likely to terminate its session at AP10 if its session time doesn't exceed 10 minutes. Users with session times exceeding 10 minutes are instead more likely to terminate their sessions at AP21. Incidentally, AP3, AP10, and AP21 are all located on the same floor of the building and are one floor below where AP14 is situated. AP6 is located one floor above AP14. In this example, a user is more likely to end its session at the lower floor as opposed to the upper floor and the network manager could use this information to infer the travel path details of the mobile user. However, a user that ends at one of the access points in the lower floor may have reached that level after

visiting some of the upper floors before going down. Hence, some caution is needed when inferring the results of $\mathbf{q_e}(t)$.

Figure 4.13 shows the probabilities for those users that have initiated their sessions at AP17 which is on the same floor and within the neighborhood of AP14. For this example, sessions with smaller durations are more likely to terminate their sessions on the upper floor at AP6. However, sessions with durations longer than 70 minutes are likely to end on the same floor at AP19 or at AP21 on the lower floor. Notice how these behaviors are quite different from those shown in Figure 4.12 primarily due to the location at which the session was initiated. The results do show some considerable fluctuations in the likelihoods as the duration of the session increases.

To assess the validity of the proposed computation for $\mathbf{q_e}(t)$, the data set obtained from [83] will again be used to examine the accuracy of the end-location predictions. A similar approach to the one taken in Chapter 3 was followed in this section. The traces recorded prior to the year 2003 were processed for computing the parameters needed to construct the expression in Equation (4.12). Only those instances that involved a transition within the same Library Building 2 were considered. The parameters were then used to generate the results for $\mathbf{q_e}(t)$ and subsequently checking each of the events in the data set that occurred on or beyond the year 2003 to see if the end-location events were accurately predicted.

A simple method was chosen to test the accuracy of the end-location predictions for a session that was initiated at access point i and lasting for t minutes, before terminating at access point j. With $\alpha = e'_i$, if the jth element in $\mathbf{q_e}(t)$ returned a higher probability than the remaining elements in the same vector, i.e. $q_{e(j)}(t) = \max\{\mathbf{q_e}(t)\}$, then the result is said to have correctly predicted the outcome. Most existing techniques would attempt to compute these end-location predictions using

the transition probabilities alone, without considering the temporal influences on the results.

Let $\mathbf{P_a}$ be an $L_A \times L_A$ transition probability matrix with elements $P_{i,j}$ that describe the probability of an arbitrary user with an active session making a transition from location i to j. Further define $\mathbf{P_0}$ as an $L_A \times L_A$ transition probability matrix with elements $P_{i,i}$ on the diagonal that describe the probability of an arbitrary user terminating its session in location i. The matrices $\mathbf{P_a}$ and $\mathbf{P_0}$ can be used to construct a simple example of computing end-location predictions using the transition probabilities alone for comparing with the proposed predictor using $\mathbf{q_e}(t)$. Both matrices were used to compute $P_{e(j)}(n)$, which is the probability of an arbitrary user in location i ending up and terminating its session in location j by the n_{th} transition. Hence, the vector $\mathbf{P_e}(n) = \left\{P_{e(j)}(n)\right\}$ can be computed as follows,

$$\mathbf{P_e}(n) \;=\; \alpha \mathbf{P_a}^{n-1} \mathbf{P_0}, \qquad \text{for} \quad n \geq 1, \tag{4.14}$$

where n in this case is the number of location transitions that a user has completed from the initial to the end location. Note that in the result for $\mathbf{P_e}(n)$ the user must have made $n-1$ location transitions, with a final transition involving the session termination. Furthermore, the results from $\mathbf{P_e}(n)$ do not provide any information on *when* a session could terminate.

Another more common method used to predict such occurrences would be to compute the origin-destination transition probability matrix, which will be denoted as $\mathbf{P_{od}}$. In this matrix the elements $P_{od_{(i,j)}}$ are the probabilities of a user's session terminating in location j given that it was initiated at location i. These probabilities do not consider the transitions that occur between the origin location i and the termination location j, if any.

Using Equations (4.13) and (4.14), as well as $\mathbf{P_{od}}$, out of the 147, 193 relevant events that were available in the data set, the overall accuracies returned by the three prediction methods were compared for the case of users having initiated their sessions at one of the 21 access points. The performance of the $\mathbf{q_e}(t)$ predictor was also examined for the case where the sojourn time behavior follows a phase-type distribution and a geometric distribution, i.e. $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and $G_{i,j}(t) \sim Geom(p)$, respectively. The elements for the $\mathbf{P_{od}}$ predictor were needed to be computed from the part of the data set that contained events logged prior to May 2003.The same values for $P_{i,j}$ and $G_{i,j}(t)$ that were computed in Section 4.2 were again used for setting up both the $\mathbf{P_e}(n)$ predictor and the $\mathbf{q_e}(t)$ predictor.

The summary of the prediction accuracy results are shown in Table 4.6 and Figure 4.14. The results show that the overall performance of the $\mathbf{q_e}(t)$ predictor achieved an accuracy of around 65% with $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and an accuracy of around 72% with $G_{i,j}(t) \sim Geom(p)$. However, they were not as high as what was computed using the conventional predictors. In a few instances, the difference in the accuracies were relatively small and the proposed end-location predictor has the added ability of determine the likelihoods of *when* (with a certain time window) a session will terminate, along with *where* the session termination will occur.

When comparing the performance of the proposed predictor between having $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$ and $G_{i,j}(t) \sim Geom(p)$, some cases (e.g., session initiations from AP13 and AP15) have shown the predictor with the phase-type distributed sojourn times to achieve a higher accuracy than those gained by assuming the sojourn times to be geometrically distributed. Other cases (e.g., session initiations from AP1 and AP3) have shown the total opposite in terms of the prediction accuracies. These findings suggest that the behavior of the sojourn times may be memoryless, but is certainly not true for all the cases shown in the results. The network manager may

again apply the proposed prediction scheme with different sojourn time distribution types depending on which one returns a higher average accuracy from the access point at which the session is initiated.

Tab. 4.6: Summary of End-Location prediction accuracy results for transitions from the access points in Library Building 2.

Session Initiations From AP	Conventional Predictors $\mathbf{P_e}(n)$ Predictor	Conventional Predictors $\mathbf{P_{od}}$ Predictor	Proposed Predictor Using MRP $\mathbf{q_e}(t)$ Predictor $G_{i,j}(t) \sim (\boldsymbol{\alpha}_{i,j}, \mathbf{S}_{i,j})$	Proposed Predictor Using MRP $\mathbf{q_e}(t)$ Predictor $G_{i,j}(t) \sim Geom(p)$	Total # of Transitions	Total # of Users
AP1	0.83867	0.8035	0.56224	0.8035	5266	1722
AP2	0.63204	0.5989	0.59792	0.59542	1122	315
AP3	0.79931	0.71646	0.5036	0.65132	7648	1977
AP4	0.76997	0.76351	0.65875	0.71387	5624	1177
AP5	0.87629	0.88096	0.88039	0.86259	2895	877
AP6	0.8154	0.81784	0.50171	0.81784	5486	1381
AP7	0.84418	0.84336	0.84336	0.84336	8670	612
AP8	0.69273	0.70527	0.42193	0.64579	1252	390
AP9	0.9437	0.93951	0.93951	0.93951	11429	1680
AP10	0.5848	0.56965	0.33028	0.47156	5080	1654
AP11	0.94214	0.93959	0.93959	0.93959	14473	2104
AP12	0.69757	0.57598	0.57598	0.60401	2420	273
AP13	0.85916	0.80998	0.80998	0.74721	5081	1199
AP14	0.79745	0.75609	0.75609	0.75609	18457	2400
AP15	0.77262	0.76732	0.71285	0.6988	7833	166
AP16	0.82039	0.79222	0.50577	0.69586	4473	862
AP17	0.77039	0.73901	0.50807	0.65152	11037	2140
AP18	0.66837	0.54625	0.59012	0.64775	818	103
AP19	0.75582	0.76608	0.75331	0.74358	12968	2593
AP20	0.67172	0.41653	0.51254	0.5305	2067	674
AP21	0.83392	0.82237	0.82237	0.82237	13094	2742
Overall	0.7803	0.7414	0.6536	0.7189		

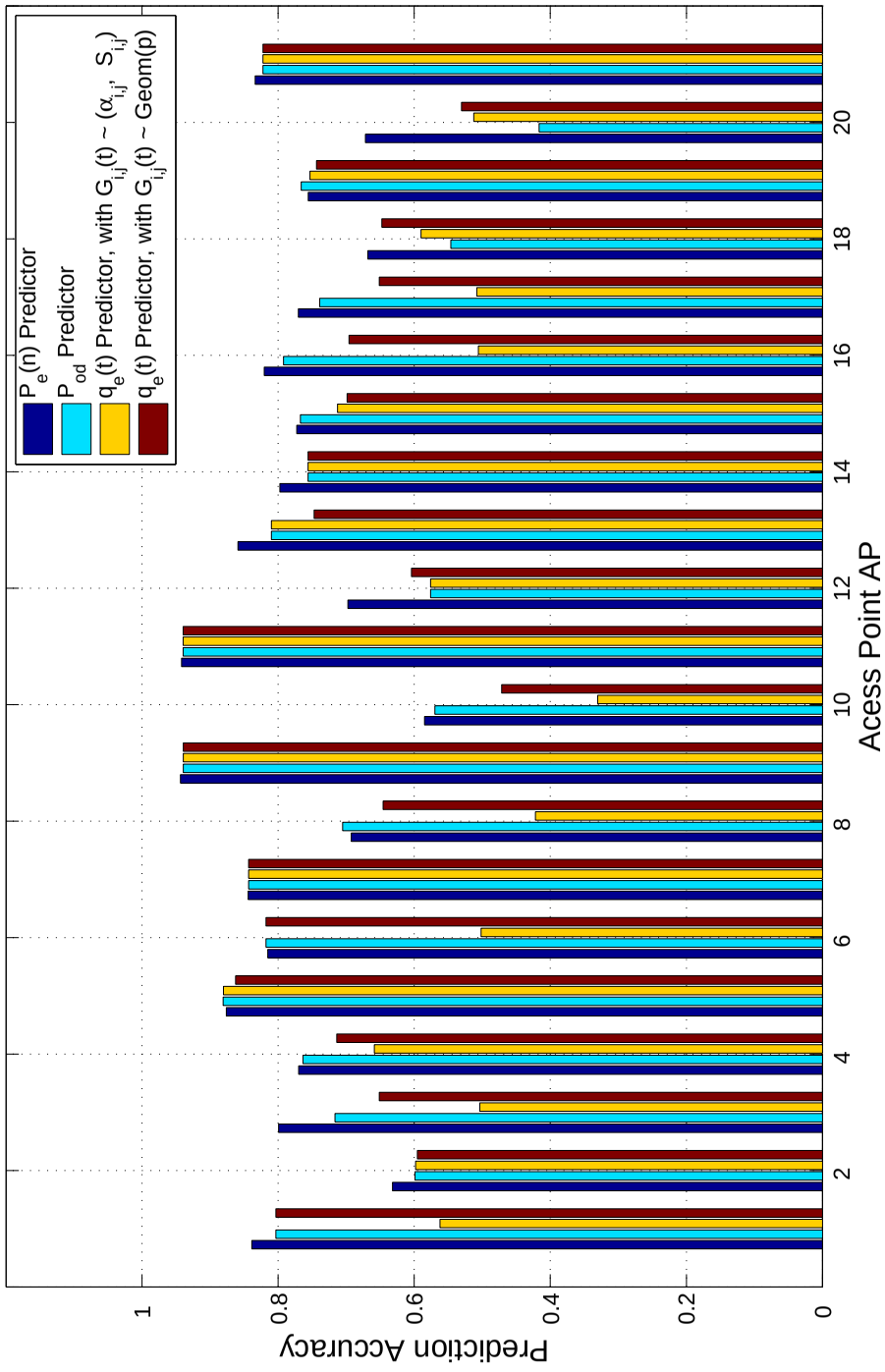


Fig. 4.14: The performance of the End-Location $\mathbf{q_e}(t)$ predictor as compared with the $\mathbf{P_e}(n)$ and $\mathbf{P_{od}}$ predictor.

5. MOBILITY & DATA TRAFFIC PREDICTION

Wireless data applications in mobile units are increasing in popularity amongst the mobile users and will likely influence the wireless traffic in future networks. E-mailing, Web-browsing, and streaming multimedia are some examples of applications that are currently in use today [1, 2]. The increasing trend in wireless data transmission will require the network providers to consider the management of the network resources that are available to the wireless data applications. Hence, the network providers will be required to safeguard the contracted QoS with the mobile user's data transmission requirements. This can be made possible by ensuring that sufficient resources are available to the users at the various locations in the network. To help with efficiently managing the network's resources for wireless data transmission, an understanding of the user's mobility relative to the data session usage patterns is needed.

Following a similar approach to the one presented in the previous chapters, this chapter will present the proposed semi-Markov model for mobility and data traffic prediction. In the previous chapters, a pair of random variables $\{X, T\}$ were defined to model both the mobility and session-activity behaviors of network users. In this chapter, an additional random variable is defined and included in the mobility model for describing the data usage by an arbitrary user. The model can ultimately be applied to predict the volume of a user's data traffic demands that are being transferred from one location to another due to mobility.

5.1 Mobility & Data Rate Prediction

In addition to assigning one random variable X to represent the location ID, a second random variable Z is defined to denote the effective transmission rate of a session between the mobile user and the network. Hence, the mobility behavior can be described by the trivariate state $\{X, Z, T\}$. The range of achievable data transmission rates can be quite large which may also depend on the type of network analyzed. Thus, it was assumed that the range can be discretized in such a way that the variable Z is an integer and represents a multiple of the data rate unit U. The parameter U is assumed to be the size (in data units per time division) of each discrete epoch and is equivalent to the minimum effective transmission rate. If R_{max} is the maximum data rate at which an arbitrary user can transmit/receive in a given network, then $\mathcal{Z} = \left\{0, 1, 2, \cdots, \lceil \frac{R_{max}}{U} \rceil \right\}$, where $Z \in \mathcal{Z}$. The manager $\lceil \cdot \rceil$ is defined as rounding up the numerical entry to the nearest integer.

The semi-Markov kernel for this bivariate model can be defined as follows,

$$Q_{(i,v)(j,w)}(t) = Pr\left\{X_{n+1} = j, Z_{n+1} = w, T_{n+1} - T_n \leq t | X_n = i, Z_n = v\right\}, \quad (5.1)$$

which is the product of the transition probability $P_{(i,v)(j,w)}$ and the sojourn time distribution $G_{(i,v)(j,w)}(t)$. Note that the elements $G_{(i,v)(j,w)}(t)$ can also depend on the size of the quantized data rate unit U. These sojourn time distributions could also be related to the number of users connecting through a particular access point, along with their transmission rates and throughput. This is especially true if we are dealing with a network that offers a shared medium to its users such as the IEEE 802.11 wireless network. Channel conditions and congestion levels can also be significant factors in determining the sojourn time distributions. Hence, one would need to take into account various conditions when computing the sojourn

time distributions for a user to change its transmission rate from vU to wU data units per time interval.

An arbitrary mobile user with a data session in location i and with a transmission rate state v can transition into location j with a change in transmission rate to state w. The probability of this transition occurring within t units of time from being in the current state is given by the element $Q_{(i,v)(j,w)}(t)$. In such a system, 3 types of transitions can occur.

- A change in location from $i \rightarrow j$ (where $i \neq j$), with or without any change in the transmission rate state (i.e. $v \rightarrow w$ with $w = v$).
- A change in the transmission rate state from $v \rightarrow w$ (where $v \neq w$) without a change in location.
- A change in session activity from being idle to active and vice versa, with an idle session having a transmission rate state of zero.

Note the above set of possible transitions assumes that a simultaneous change in both location and transmission rate state (with $v > 0$) is permitted. This is particularly true when dealing with networks that offer a shared medium to its users, such as IEEE 802.11 WLANs. In such networks, the effective data-rate that is experienced by a user depends on the number of other users that are utilizing the same shared medium that is managed by a particular access point.

Consider a network with L different locations that are each served by a single access point, where $L < \infty$. Hence, the random variable $X \in \mathcal{X}$ defines the user's location in the network where the set $\mathcal{X} = \{1, 2, \cdots, L\}$. Furthermore, let $S = \lceil \frac{R_{max}}{U} \rceil$, such that $\mathcal{Z} = \{0, 1, 2, \cdots, S\}$, where $Z \in \mathcal{Z}$. Therefore, the semi-Markov kernel matrix $\mathbf{Q}(t)$ for this model will have the following form.

$$\mathbf{Q}(t) = \begin{pmatrix} Q_{1,1}(t) & Q_{1,2}(t) & \cdots & Q_{1,L}(t) \\ Q_{2,1}(t) & Q_{2,2}(t) & \cdots & Q_{2,L}(t) \\ \vdots & \vdots & \ddots & \ddots \\ Q_{L,1}(t) & Q_{L,2}(t) & \cdots & Q_{L,L}(t) \end{pmatrix}, \qquad \text{where} \tag{5.2}$$

$$Q_{i,j}(t) = \begin{pmatrix} Q_{(i,0)(j,0)}(t) & 0 & \cdots & 0 \\ 0 & Q_{(i,1)(j,1)}(t) & \cdots & Q_{(i,1)(j,S)}(t) \\ \vdots & & \ddots & \\ 0 & Q_{(i,S)(j,1)}(t) & \cdots & Q_{(i,S)(j,S)}(t) \end{pmatrix} \quad \text{for} \;\; i \neq j, \tag{5.3}$$

$$Q_{i,i}(t) = \begin{pmatrix} 0 & Q_{(i,0)(i,1)}(t) & Q_{(i,0)(i,2)}(t) & \cdots & Q_{(i,0)(i,S)}(t) \\ Q_{(i,1)(i,0)}(t) & 0 & Q_{(i,1)(i,2)}(t) & \cdots & Q_{(i,1)(i,S)}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q_{(i,S)(i,0)}(t) & Q_{(i,S)(i,1)}(t) & \cdots & Q_{(i,S)(i,S-1)}(t) & 0 \end{pmatrix}. \tag{5.4}$$

The transitions in $Q_{i,j}(t)$, for $i \neq j$ include those that involve a change in location due to mobility, which may also include a change in the transmission rate state. The elements in $Q_{i,i}(t)$ comprise the transitions from one transmission rate state to another without a change in location, including the transitions in switching from idle to active. A similar definition can be given for computing the elements $q_{(i,v)(j,w)}(t)$ which describe the transitions occurring *at* time t rather than within time t. Notice that the idle $\leftrightarrow$ active transitions of a user's session are assumed to occur without a simultaneous change in location, which is the reason for the zeros across the first row and the first column of the matrix given in Equation (5.3).

The case of $S = 1$ reverts this model back to the original model described in the previous chapters, thus depicting those users with a single service-type, e.g. voice calls. Moreover, the accuracy of any prediction computed using this model can be significantly influenced by the chosen size of the quantized transmission rate levels U. The extended applications proposed in Chapter 4 can be similarly adapted to the semi-Markov mobility model given in this section.

There may be instances where the maximum achievable data rate $(R_{max})_i$ depends on the access point that is serving the location i. This is true for cases where different access points are provisioned with dissimilar bandwidth allocations. Hence, the state Z in the same model can be re-defined such that the set $\mathcal{Z}_i = \{0, 1, 2, \cdots, S_i\}$, where $S_i = \lceil \frac{(R_{max})_i}{U} \rceil$, for $Z \in \mathcal{Z}_i$, and is dependent on the state $X = i$. This heterogeneous set of transmission rate states allow for a more general model. For simplicity, the remainder of this chapter will assume the homogenous case.

The model proposed in this section assumes the availability of information related to the changes in the data-rate and locations of a mobile user. May of the traffic traces that were found available online had only recorded either of the two types of information alone but never together. Some do not explicitly keep track of the changes in the effective transmission rates for each wireless terminal and that such information would need to be derived from the traffic data, if possible. Other traces were instead found to keep a temporal record of a user's data traffic in terms of the volume of bytes/packets exchanged between the mobile terminal and the access point. This other form of information may also be useful with modeling the mobility and data traffic behavior of an arbitrary user, as shown in the next section.

5.2 Mobility & Data Volume Prediction

In a few of the traces that were available online to the research community, the user's data traffic logs included information on the number of bytes transmitted and received per session. An example of such traces is given in [89]. The contents from such logs are insufficient to compute the changes in the effective transmission rates that are needed to form the mobility model defined in the previous section. In this section, another mobility model is proposed for the case where the information on the size of the data traffic (rather than the changes in the transmission rates) is available in the user's logs. This mobility model could also be used for predicting the changes in the wireless traffic in terms of volumes of data units transferred between the network and the mobile user.

A similar approach to the one adopted in the previous section will be used to define the mobility model given in this section. This mobility model will again assume a trivariate state $\left\{\check{X}, \check{Z}, T\right\}$, with the random variable $\check{X}$ having the same definition as X given in the previous section. The random variable $\check{Z}$ is assumed to denote the amount of data that has completed its transfer between the user and the network. A data session between the network and the mobile user can involve the transfer of a large number of bytes/packets per session. Let D_{max} be the maximum number of bytes that can be transferred between a user and the network, per session. We further assume that the amount of data transferred can be discretized into batches of size B bytes. Therefore, the maximum number of batches that can be transferred is $\check{S} = \lceil \frac{D_{max}}{B} \rceil$.

For predicting the increasing changes in the amount of data units transmitted per user, as well as changes in the user's location due to mobility, we define the

semi-Markov kernel $\check{Q}_{(i,v)(j,w)}(t)$ as follows.

$$\check{Q}_{(i,v)(j,w)}(t) \;=\; Pr\left\{\check{X}_{n+1}=j, \check{Z}_{n+1}=w, T_{n+1}-T_n \leq t | \check{X}_n=i, \check{Z}_n=v\right\}, \quad (5.5)$$

where $\check{X} \in \check{\mathcal{X}}$ such that $\check{\mathcal{X}} = \{1,2,\cdots,L\}$, with $L < \infty$, and $\check{Z} \in \check{\mathcal{Z}}$ such that $\check{\mathcal{Z}} = \left\{0,1,2,\cdots,\check{S}\right\}$. Note that $\check{\mathcal{Z}} = 0$ describes the state of a user that is not engaged in an active session with the network, i.e. the user's session is idle. $\check{Q}_{(i,v)(j,w)}(t)$ is defined as the probability of an arbitrary user making a transition from location i to location j, while having completed the transmission of $(w-v)$ batches of data within a time t, where $w \geq v$. A prediction scheme that utilizes the transition probabilities given by this model looks at determining the likelihood of finding an arbitrary user in location j, having also transmitted w batches of data (i.e., wB bytes) within a time interval t relative to the current time. This prediction is made given that the arbitrary user at the current time is being served in location i and has already completed the transmission of v batches of data (i.e., vB bytes). In this particular mobility model, the following are the set of possible transitions.

- A change in location while the user's network session is idle without transmitting any data, i.e. $v = w = 0$ and $i \neq j$.
- A change in location without the transmission of any data from a user's ongoing active session (no transmission in the interval t), i.e. $v = w$ and $i \neq j$.
- A change in location with the transmission of $(w-v)$ batches of data by a user's ongoing active session in the current location i, i.e. $w > v$ and $i \neq j$.
- The transmission of $(w-v)$ batches of data with the user terminating its current session in the same location, i.e. $w > v$ with $i = j$. Note that a total of wB bytes have been transmitted in the user's session.

The semi-Markov kernel matrix $\check{\mathbf{Q}}(t)$ for this mobility model can be written as follows.

$$\check{\mathbf{Q}}(t) = \begin{pmatrix} \check{Q}_{1,1}(t) & \check{Q}_{1,2}(t) & \cdots & \check{Q}_{1,L}(t) \\ \check{Q}_{2,1}(t) & \check{Q}_{2,2}(t) & \cdots & \check{Q}_{2,L}(t) \\ \vdots & \vdots & \ddots & \ddots \\ \check{Q}_{L,1}(t) & \check{Q}_{L,2}(t) & \cdots & \check{Q}_{L,L}(t) \end{pmatrix}, \tag{5.6}$$

where, for $i \neq j$,

$$\check{Q}_{i,j}(t) = \begin{pmatrix} \check{Q}_{(i,0)(j,0)}(t) & \check{Q}_{(i,0)(j,1)}(t) & \check{Q}_{(i,0)(j,2)}(t) & \cdots & \check{Q}_{(i,0)(j,\check{S})}(t) \\ 0 & \check{Q}_{(i,1)(j,1)}(t) & \check{Q}_{(i,1)(j,2)}(t) & \cdots & \check{Q}_{(i,1)(j,\check{S})}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \check{Q}_{(i,\check{S}-1)(j,\check{S}-1)}(t) & \check{Q}_{(i,\check{S}-1)(j,\check{S})}(t) \\ 0 & 0 & 0 & \cdots & \check{Q}_{(i,\check{S})(j,\check{S})}(t) \end{pmatrix}, \tag{5.7}$$

and,

$$\check{Q}_{i,i}(t) = \begin{pmatrix} 0 & \check{Q}_{(i,0)(i,1)}(t) & \check{Q}_{(i,0)(i,2)}(t) & \cdots & \check{Q}_{(i,0)(i,\check{S})}(t) \\ 0 & 0 & \check{Q}_{(i,1)(i,2)}(t) & \cdots & \check{Q}_{(i,1)(i,\check{S})}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \check{Q}_{(i,\check{S}-1)(i,\check{S})}(t) \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{5.8}$$

The probabilities given by the elements in $\check{Q}_{i,j}(t)$ evaluate the likelihood of a user changing its association from the access point serving location i to the one serving location j due to mobility, for $i \neq j$. These transitions may also involve the successful transfer of $w - v$ batches of data (with $w > 0$) before completing the location change. This includes the case of $w = v$ where no data is transmitted by the user during the time interval t. Idle users in terms of network sessions are

also accounted for in the $\check{Q}_{i,j}(t)$ elements with $v = 0$. Such users may also continue to remain idle after making a location transition, as given by the elements with $w = 0$. The probabilities given by the elements in $\check{Q}_{i,i}(t)$ evaluate the likelihood of a user completing its network session without making a location transition, while also having successfully transmitted $(w - v)$ batches of data. Note that $\check{Q}_{(i,0)(i,0)}(t) = 0$ (as given by the upper left-most element in matrix $\check{Q}_{i,i}(t)$) since that transition does not represent a "renewal" in the process.

In addition to making future location predictions, the probabilities $\check{Q}_{(i,v)(j,w)}(t)$ may also be applied for predicting how long the network is expected to sustain a user's ongoing active session through a particular access point These predictions could also assist with evaluating the expected volume of data that is to be transferred between an arbitrary user and an access point during a particular time interval. Therefore, such information could yield a spatial-temporal prediction of the data traffic behavior within a network. The prediction results may also help with identifying the expected amount of workload (in terms of data bytes to be transferred) that an arbitrary user brings into each location. Note again that the strength of the predictions may also depend on the choice of the data batch size B used to discretize the range of transferred data bytes $[0, D_{max}]$.

5.3 Numerical Example

During the course of this research, it was not possible to obtain a set of traffic that would contain the information needed to apply the proposed mobility model presented in this chapter for validation and assessment purposes. Much of the traces found had not included both the location transitions and the changes in the data usage information in their traffic logs. There were some traces (e.g., [89]) that provided partial details on the amount of traffic that had flowed between the mobile users and the network but they did not include the changes in the traffic flow as the user moves from one access point to the next. Combining such traces with other independent data sets that include the user movement patterns is one option that could be explored. However, combining two or more independent traffic traces may result in an unrealistic analysis since the behaviors of the disparate traces are unrelated to each other.

It would have been preferable to examine the use of this model in a manner that is similar to what was done in the previous chapters. Due to the lack of real data, a numerical example will instead be given. The purpose of the example is to show how the results from the mobility model proposed in this chapter can be interpreted for the benefit of making predictions. Even though the example given in this section is restricted to applying the model proposed in this chapter, the prediction methods and further applications described in the previous chapters can also be employed for this data network mobility model.

To supply the synthetic data needed to construct this mobility model, a networking environment was simulated with users roaming between numerous access points within some coverage areas, while randomly engaging in a data session at various locations. The node mobility simulator utilized in Section 4.4, which was

developed and based on the Smooth Mobility model proposed by Bettstetter in [93], was re-used again in this section. The same parameters given in Table 4.4 were also assigned in the simulator for the nodal mobility behavior in the network. An IEEE 802.11 type of wireless network was chosen as the setting for the simulator, with each access point supervising its wireless channel that is shared amongst the users that are associated with it.

The simulator was extended to involve the initiation of a data session between a mobile user and the network. Several researchers, such as Crovella and Bestavros in [99] and Fraleigh et al. [100], have identified the self-similar nature of the data traffic behavior. They have found it best to model these types of traffic using heavy-tailed distributions such as Pareto and Weibull. The parameters that define these traffic behaviors can be influenced by numerous factors including the network protocol, application, and network terminal type that it engaged by the users. Incorporating such behaviors into the simulation can yield a more realistic response. However, they could also introduce further complexities which may require the extra tuning of their parameters to achieve some level of rational behavior. To make matters simple, it was elected to model the transmission time and the session idle time of a user as a random process that follows a geometric distribution. This is similar to what was assumed in the very early works on traffic modeling. In this section, the interest is not in simulating a lifelike mobility model and is instead primarily focused on providing an adequate example and demonstrate the applicability of its results for prediction purposes.

In the simulation, each location is assumed to manage the same quota of shared medium/resources. An active session receives a portion of these resources depending on how many users with active sessions are associated with the same access point. For example, a session would be granted the entire bandwidth and running at the

maximum permissible transmission rate if no other sessions are running within the same coverage area of the access point. This session's transmission rate would reduce as the number of other active sessions requiring the usage of the same shared medium increases. The time taken to complete a user's active session at some maximum permitted transmission rate is assumed to be geometrically distributed with $p = 0.01$. The remaining session times are adjusted throughout the simulation depending on the number of users with active sessions that are associated with the same access point. These session times will have a mean of $1/p = 100$ seconds. This value for the mean session time was chosen to ensure that an ample number of sessions will likely remain active while being involved in a location transition. This made it possible to collect enough data to compile the information needed to construct the mobility model. The time a user's session remains idle is also assumed to be geometrically distributed with $p = 0.002$.

For this example, the model proposed in Section 5.1.1 was used which considers the transitions involving both the changes in location and transmission rates. The elements $Q_{(i,v)(j,w)}(t)$ for this model were evaluated from the mobility and traffic data logs accumulated by the simulation. The logs incorporated the details of each mobile user's location progressions, as well as the times at which the user was running its session at one of the 5 discrete transmission rate levels, with level 5 being the maximum transmission rate. The session idle times were also recorded in the data and was denoted in the model as having a transmission rate level of 0. Hence, in this example a mobile user can be in any of the 16 locations operating at any of the 6 different transmission rate levels. Consequently, this leads to a total number of 96 different states for this mobility model. This example also illustrates the possibility of ending up having a model with a relatively large state space when dealing with mobility and data traffic predictions in practical situations. In practice, the location

approximation given in Section 3.3 could also be applied in such circumstances for reducing the state space.

For the elements $Q_{(i,v)(j,w)}(t)$ in this example, the transition probabilities were computed in the usual manner. The conditional state sojourn time distributions $G_{(i,v)(j,w)}(t)$ were evaluated by selecting a distribution that closely fits the associated sojourn time data. This was accomplished by the distribution fitting tool in Matlab. The log likelihood results that were returned by the fitting tool was used to judge the choice of the best distribution to be assigned for each of the the model's elements. The results in this example yielded the log-normal distribution as being the best-fitting distribution for the sojourn times involving location transitions with active sessions at various transmission rate levels. The Weibull distribution was more appropriate at modeling the sojourn times for transitions with idle sessions (i.e., data-rate level of 0), based on the higher log-likelihood result returned using Matlab's distribution fitting toolbox. This feature of designating more than one type of distribution for modeling the state sojourn time behaviors demonstrates one of the key strengths in adopting the proposed method.

After running the simulation to generate the traffic data needed to compute the MRP elements in the example, the behavior of some of these elements were examined. A subset of them have been selected for analyzing the several inferences that can be made from such results. These discussions will also include how the results could be used for prediction purposes. The numerical results given next review how an arbitrary mobile user situated in location 5 is expected to behave in this simulated network. Hence, the focus is on understanding the future progressions of a user in location 5. A similar analysis can be performed on those users that are situated at the other locations.

For users that have just entered location 5 with idle sessions (i.e., data-rate level

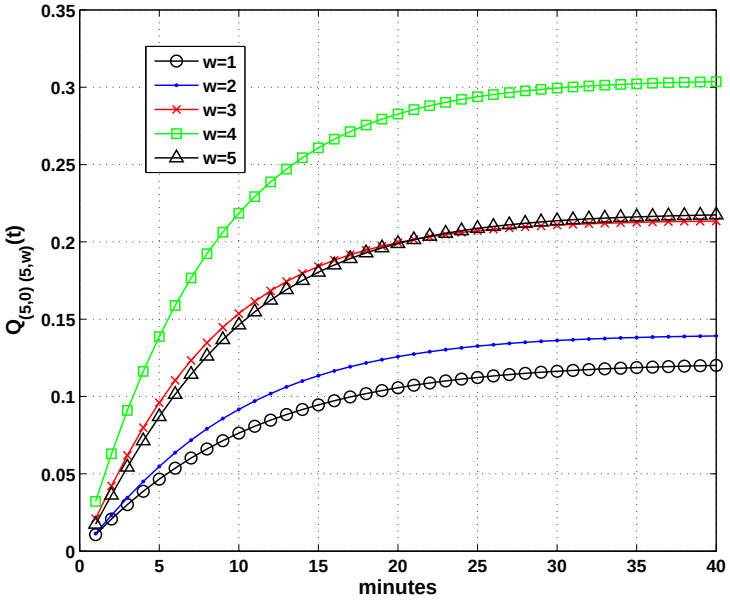


Fig. 5.1: A plot of the probabilities $Q_{(i,v)(j,w)}(t)$ for a user initiating a session in location 5.

of 0), the likelihoods of them initiating a session with the network within a certain time period is given by the plots shown in Figure 5.1. These plots also illustrate how the possibilities of having these sessions initiated at one of the 5 different transmission rate levels vary with time, with level 5 amounting to the highest transmission rate that can be offered by the access point at the given location. Out of the 5 possible levels, the model seems to indicate that if a user in location 5 is expected to launch a session with the network then it will most likely do so with a level 4 transmission rate, as indicated by results for $Q_{(5,0)(5,4)}(t)$. This suggests that users in the given location have a good chance at having their sessions initiated with a transmission rate that is close to the maximum. However, this does not imply that their sessions will continue to remain active at the same transmission rate level. It may even change for the better or worse and the chances of such transitions being governed by the other elements in the model. This information could further be

used to predict the traffic load at the given location. Since the simulated network example assumes that the access point's bandwidth is shared amongst its users, a lower transmission rate would be due to a higher number of users that are simultaneously accessing the network. If on the other hand the probabilities in Figure 5.1 were higher for the lower transmission rate levels, then the network load at the given location is expected to be quite high. This information could prompt the network managers to consider taking certain measures and actions at the given access point, e.g. increase the bandwidth. Such actions could also include communicating the option of accepting a downgraded performance to the user's running application (e.g. running a streaming video at a lower resolution), or even exploring the possibility of having to "borrow" some bandwidth from neighboring sources for sustaining the high traffic loads.

The next top-most probabilities shown in Figure 5.1 predict the possibility of having a user in location 5 initiating a session with transmission rate levels of 3 or 5. The odds of being granted either of the two data-rate levels are quite close together where one surpasses the other depending on how long the user has waited before initiating its session. The results suggest that the longer the user's terminal remains idle in that location, the better the chances it has at accessing the network with the highest possible transmission rate, as shown by the results for $Q_{(5,0)(5,5)}(t)$. Depending on how significant the difference is in the transmission rate levels and the user's application demands, the network could opt in such circumstances to delay the user's access to the network if immediately required. The decision would be taken if it would improve the chance of granting the user a higher transmission rate level, after delaying the initial access by some tolerable amount of time. Making these decisions could become more plausible if the difference in the data-rate levels are much wider at such circumstances with close and temporally inter-changing

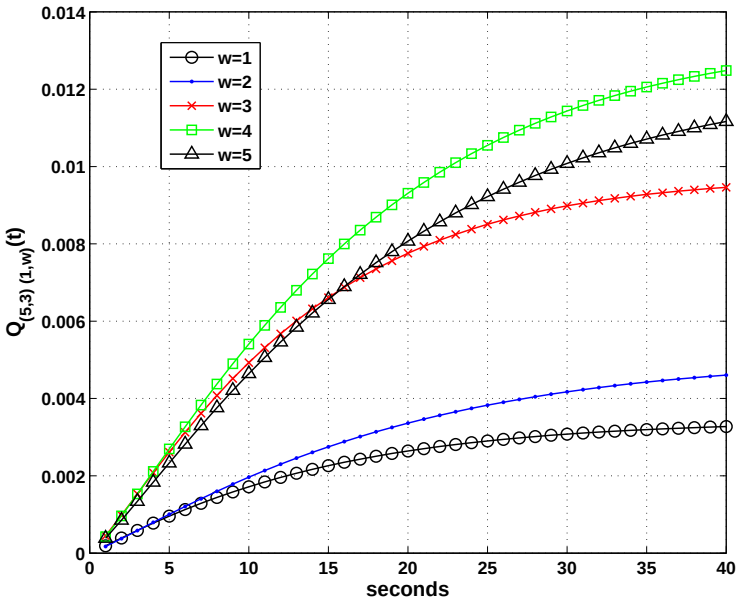


Fig. 5.2: A plot of the probabilities $Q_{(i,v)(j,w)}(t)$ for a user transitioning into location 1.

likelihoods. Such decisions could assist the network with being more proactive with managing its resources to maximize the gross performance perceived by its users.

Figures 5.2 and 5.3 illustrate the possibilities of a user transitioning from location 5 into the neighboring locations 1 and 6, while running a session at transmission rate level 3, respectively. These transitions include the possibility of having a user's transmission rate changing to some other level due to how busy the access point is at the new location. Note that the time scales in both figures are in seconds whereas in Figure 5.1 it was in minutes. The difference was attributed to the simulation assuming that the time a mobile user spends accessing the network is much less when compared with the time its session is idle. On the whole, if the network was to predict that a user in such a state will end up in one of these two locations within a certain time period, then there is a higher chance that its transmission rate can either remain the same or even be upgraded at the new locations. A user

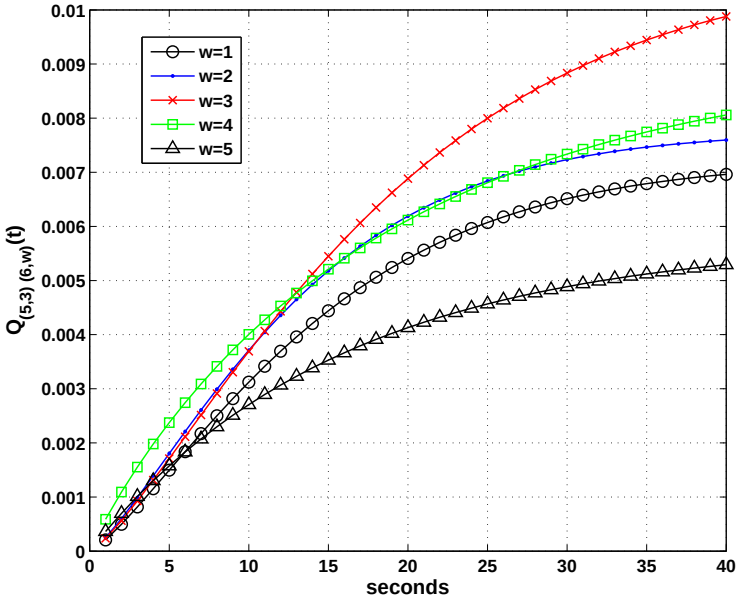


Fig. 5.3: A plot of the probabilities $Q_{(i,v)(j,w)}(t)$ for a user transitioning into location 6.

will unlikely achieve the highest transmission rate level after making a transition into location 6, based on the low probabilities depicted by $Q_{(5,3)(6,5)}(t)$ in Figure 5.3. However, let us assume the circumstance where the reverse was true such that likelihoods of having to reduce the transmission rate levels at the new locations were much higher than the rest. In this instance, the network could attempt to negotiate a downgrade in performance with the user's running application in anticipation of the expected drop in transmission rate. Other prevention schemes for avoiding sudden-disruptions could also be invoked depending on how the network managers choose to manage their own resources.

Out of all the possible outcomes for a user running a session with a transmission rate level of 3 in location 5, the five topmost likely transitions are given in Figure 5.4. The results indicate that it is highly probable for a user in such a state to make a transition into the neighboring location 9, but a reduction in the transmis-

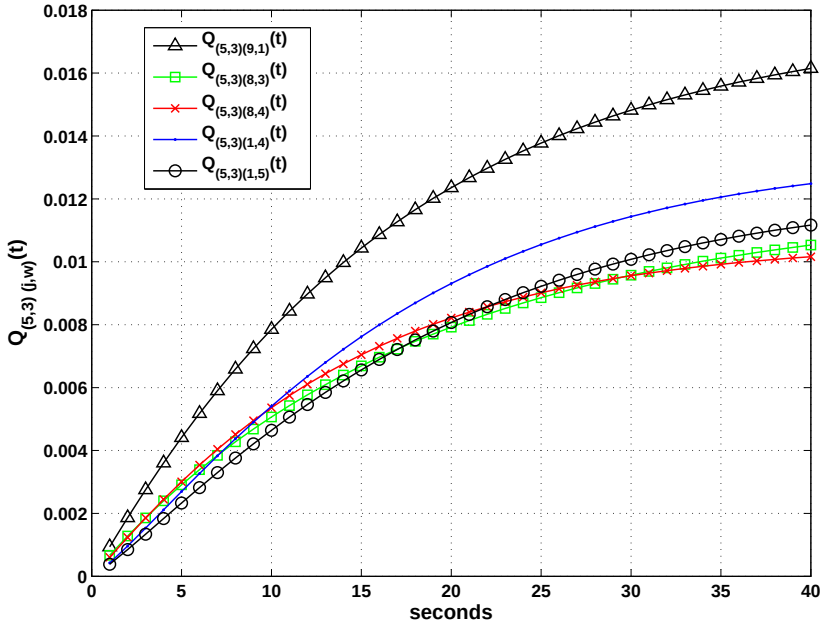


Fig. 5.4: A plot of the topmost likely outcomes for a user in location 5 running a session at a transmission rate level of 3.

sion rate level from 3 to 1 in the new location is also to be expected. This most likely outcome suggests a drop in the network performance perceived by the user. However, the probabilities of the remaining outcomes in Figure 5.4 display a more promising performance in terms of the transmission rate levels to be achieved by the user. This might encourage the network to be more focused in its concern with the transition of users into location 9 (from location 5) and initiate some course of action that might help with moderating and/or preventing the drop in performance.

The example analyzed in this section has demonstrated the benefits of exploiting the likelihoods returned by the proposed MRP-based mobility and data rate model. The model can be applied for predicting the changes in both a user's location and session transmission rate. Not only can such information be used to anticipate the usage of the network's resources at various location sites, but it can also provide

some insight on the traffic volume loads experienced by the access points. The network could take advantage of these predictions and proactively administer some performance disruption prevention scheme when needed. This could assist with better managing the overall resources as well as attempting to maintain a certain level of performance for both the users and the network.

6. CONCLUSIONS AND FUTURE WORK

6.1 Contribution and Comments

In this thesis, a wireless user mobility model was proposed and was set up as a Markov Renewal process. The mobility model was mainly developed for the purpose of predicting the subsequent transitions in mobility and activity of network users within a time period t. The proposed model can be applied for making the necessary predictions for both single and multiple transitions. A method for applying the model to generate end-location predictions was also shown. The model can be used by network managers for the purpose of efficient network resource management as well as ensuring a certain level of QoS perceived by the mobile users. Huang et al. in [9] and François Leduc in [10] have shown how a network's admission control that is combined with a mobility prediction scheme can improve the improve the overall performance perceived by both the mobile users and the network.

The mobility model was also extended to deal with predicting the mobility of users with *data* traffic. The proposed model is suitable for predicting user mobility and the length of time at which they are active and considers network architectures that support both voice and data connectivity. The latter type of network service is likely to dominate the network usage in the upcoming years. Hence, future wireless networks will likely require an understanding of how the amount of traffic is transferred between the different locations.

The data set that was obtained from the CRAWDAD repository [83] was in-

strumental with validating and showing the benefits of utilizing the proposed semi-Markov mobility model. The results given in this thesis have shown the improvements in the prediction accuracy that can be achieved by applying the proposed semi-Markov prediction scheme. The performance of the proposed scheme was compared with some of the conventional Markov predictors and was shown to yield a higher accuracy on the overall. These improvements were due to the inclusion of the temporal influences in the predictions. Furthermore, the results demonstrated one of the key strengths of the proposed model in assuming any general distribution for the user sojourn times.

Other common schemes apply machine learning techniques and neural networks [26, 84] for developing the mobility predictors and rely heavily on the availability of data. Modeling the mobility behavior using neural networks have shown to offer an improvement on the prediction accuracy when compared with other methods [84]. However, such schemes can be rather complex to build and the training process is generally slower than other schemes. Markov predictors are favored by many researchers due to their fast and simple training process as well as their generality and domain independence [85]. The proposed semi-Markov model expands on the Markov predictors and having the same advantages, along with including the temporal influences in the predictions.

The work presented in this thesis had been published in part in [101] and [102].

6.2 Model Limitations

One general disadvantage of employing Markov (and the same for semi-Markov) predictors is their slow re-learning capability [84]. In other words, it might require a large amount of data to update the set of probabilities in the model in order to reflect certain significant changes in the user behavior. In such situations, the Markov predictor may generate quite a large number of inaccurate predictions during the length of time it takes to re-learn and update the new behavior.

The proposed mobility model assumes the availability of the traffic data needed to compute its parameters for generating the needed predictions. This makes the model applicable for networks that have been active for a considerable length of time and would have enough data collected for constructing the semi-Markov model. Applying such a model for and during an initial network setup may still be possible if some data from a similar network can be used to construct and utilize the model during the initial phase of the network commissioning. The model further assumes that the behavior found in the data is stationary, which is a common assumption amongst all Markov predictors.

The semi-Markov mobility model is limited to describing the changes in the behavior of the user in terms of his/her location from the network's perspective, the current activity with the network (e.g., active, idle, transmission rate), and the time window during which the changes in the events occur. The proposed model assumes the behavior of each user to be independent and have no influence on the the other users within the surrounding population.

6.3 Future Extensions

To improve on the validity of our mobility prediction, other sets of traffic traces would need to be investigated and preferably ones that cover other types of network architectures, such as cellular and vehicular wireless networks. A considerably large repository would also be helpful at examining whether segregating the traffic traces based on seasonality can improve the prediction results. The filtering-out of infrequent and uncommon events from the traffic traces to better construct the mobility model would also need to be reviewed.

The proposed scheme is limited to modeling the behavior of individual and independent users. The work can be extended to include group behavior modeling (similar to what was proposed by Zhou et al. in [55]) as well as the influence these users may have on the others within their group. Other relevant information, such as road topology [38] and geographical constraints and/or landmarks [40], can also be incorporated into the model. The improvements gained by including such additional information in the proposed prediction scheme would also need to be examined.

Decision-making under uncertainty plays a crucial role in our proposed mobility prediction scheme and an efficient way of making such decisions is yet to be investigated. The errors in the prediction results could also be exploited for the possibility of improving future predictions. Hence, the inclusion of a feedback mechanism into the prediction scheme could also be examined. Though it was briefly mentioned in some of the sections, another area that is worth pursuing further in this work is to extend the proposed prediction scheme to handle multiple classes of users and/or applications, as well as looking into the decision-making that is involved in such cases.

REFERENCES

[1] A. Kivi, "Mobile Data Service Usage Measurements - Results 2005-2006", *COIN National Project - Helsinki University of Technology*, 2007 Project Report, http://www.netlab.hut.fi/tutkimus/coin/.

[2] A. Kivi, "Mobile Data Service Usage Measurements - Results 2005-2007", *COIN National Project - Helsinki University of Technology*, 2008 Project Report, http://www.netlab.hut.fi/tutkimus/coin/.

[3] T. S. Rappaport, "Wireless Communications: Principles and Practice", *Prentice Hall PTR*, 2002.

[4] M. Schwartz, "Mobile Wireless Communications", *Cambridge University Press*, 2005

[5] N. D. Tripathi, J. H. Reed, and H. F. VanLandinoham, "Handoff in Cellular Cystems", *IEEE Personal Communications*, Vol.5, issue 6, December 1998.

[6] W. Soh and H. S. Kim, "QoS Provisioning in Cellular Networks Based on Mobility Prediction Techniques", *IEEE Communications Magazine*, January 2003.

[7] D. Hong and S. S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures", *IEEE Transactions in Vehicular Technology*, Vol. 35, No. 3, August 1986.

[8] M. S. Chiu and M. A. Bassiouni. "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning", *IEEE Journal on Selected Areas in Communications*, 18(3), March 2000.

[9] Q. Huang, S. Chan and M. Zukerman, "Improving Handoff QoS With or Without Mobility Prediction", *Electronics Letters*, Vol. 43, No. 9, April 2007.

[10] J. François and G. Leduc, "Mobility Prediction's Influence on QoS in Wireless Networks: A Study on a Call Admission Algorithm", *Symposium on Modeling and Optimization in Mobile, AdHoc, and Wireless Networks*, WIOPT 2005.

[11] J. Gossa, A. G. Janecek, K. A. Hummel, W. N. Gansterer, and J. Pierson, "Proactive Replica Placement Using Mobility Prediction", *In IEEE 9th International Conference on Mobile Data Management Workshops*, MDMW 2008.

[12] A. Jayasuriya, "Handover Channel Allocation Based on Mobility Predictions", *Advanced Wired and Wireless Networks*, Springer 2005.

[13] P. N. Pathirana, A. V. Savkin and S. Jha, "Location Estimation and Trajectory Prediction for Cellular Networks with Mobile Base Stations", *IEEE Transactions on Vehicular Technology*, Volume 53, Issue 6, November 2004.

[14] W. Su, S. Lee, and M. Gerla, "Mobility Prediction in Wireless Networks", *MILCOM 2000. 21st Century Military Communications Conference Proceedings*, Volume 1, 22-25 October 2000.

[15] R. Chellappa-Doss, A. Jennings and N. Shenoy, "User Mobility Prediction in Hybrid and Ad Hoc Wireless Networks", *In Australian Telecommunications Networks and Applications Conference*, 2003.

[16] R. Chellappa, A. Jennings and N. Shenoy, "A Comparative Study of Mobility Prediction in Fixed Wireless Networks and Mobile Ad Hoc Networks", *IEEE International Conference on Communications*, May 2003.

[17] H.-W. Ferng, W.-Y. Kao, D. Shiung, C.-L Liu, H.-Y. Chen, and H.-Y. Gu, "A Dynamic Resource Reservation Scheme with Mobility Prediction for Wireless Multimedia Networks", In *IEEE 60th Vehicular Technology Conference VTC2004-Fall*, September 2004.

[18] F. De Rango, P. Fazio and S. Marano, "Mobility Prediction and Resource Reservation in WLAN Networks Under a 2D Mobility Model", *IEEE 64th Vehicular Technology Conference*, Fall 2006.

[19] A. Jayasuriya and J. Asenstorfer, "Mobility Prediction Model For Cellular Networks Based on the Observed Traffic Patterns", *Proceedings of Wireless and Optical Communications*, WOC, July 2002.

[20] J. Chan, S. Zhou and A. Seneviratne, "A QoS Adaptive Mobility Prediction Scheme for Wireless Networks", *IEEE GLOBECOM*, November 1998.

[21] J. Chan and A. Seneviratne, "A Practical User Mobility Prediction Algorithm for Supporting Adaptive QoS in Wireless Networks", In *Proceedings IEEE International Conference on Networks*, October 1999.

[22] S. Choi and K. G. Shin, "Adaptive Bandwidth Reservation and Admission Control in QoS-Sensitive Cellular Networks", *IEEE Transactions on Parallel and Distributed Systems*, 13(9), September 2002.

[23] S. Kwon, H. Park and K. Lee, "A Novel Mobility Prediction Algorithm Based on User Movement History in Wireless Networks", In *Systems Modeling and Simulation: Theory and Applications: Third Asian Simulation Conference, AsianSim 2004*, October 2004.

[24] S. Michaelis and C. Wietfeld, "Comparison of User Mobility Pattern Prediction Algorithms to Increase Handover Trigger Accuracy", *IEEE 63rd Vehicular Technology Conference*, Spring 2006.

[25] F. Erbas et al., "On the User Profiles and the Prediction of User Movements in Wireless Networks", *The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Volume 5, September 2002.

[26] J. Capka and R. Boutaba. "Mobility Prediction in Wireless Networks Using Neural Networks", *Lecture Notes in Computer Science*, Vol. 3271, Springer 2004.

[27] P. Prasad and P. Agrawal, "Mobility Prediction for Wireless Network Resource Management", *In 41st Symposium on System Theory*, March 2009.

[28] L. Song, U. Deshpande, U. C. Kozat, D. Kotz, and R. Jain. "Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning", In IEEE INFOCOM 2006

[29] M. H. Sun and D. M. Blough, "Mobility Prediction Using Future Knowledge", *Proceedings of the 10th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, October 2007.

[30] F. Yu and V. Leung, "Mobility-based Predictive Call Admission Control and Bandwidth Reservation in Wireless Cellular Networks", *Computer Networks*, Vol. 38, 2002.

[31] A. Bhattacharya, S. K. Das and S. Roy, "Towards a Universal Model for Personal Mobility Management", *Wireless Communications and Networking Conference*, Volume 3, September 2000.

[32] L. Song, D. Kotz, R. Jain and X. He, "Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data", *IEEE Transactions on Mobile Computing*, Vol. 5, No. 12, December 2006.

[33] M. Park, J. Hong, and S. Cho, "Two-Stage User Mobility Modeling for Intention Prediction for Location-Based Services", *IDEAL - Lecture Notes in Computer Science*, Springer 2006.

[34] I. F. Akyildiz and W. Wang, "The Predictive User Mobility Profile Framework for Wireless Multimedia Networks", *IEEE/ACM Transactions on Networking*, Vol. 12, No.6, December 2004.

[35] D. A. Levine, I. F. Akyildiz and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks using the Shadow Cluster Concept", *IEEE/ACM Transactions on Networking*, 5(1), February 1997.

[36] T. Liu, P. Bahl and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks", *IEEE Journal on Selected Areas in Communications*, 16(6), August 1998.

[37] W. Wang, Y. Cao, D. Li, and Z. Qin, "Markov-Based Hierarchical User Mobility Model", *Proceedings of the IEEE International Conference on Wireless and Mobile Communications*, ICWCMC 2007.

[38] W.-S. Soh and H. S. Kim, "Dynamic Bandwidth Reservation in Cellular Networks using Road Topology Based Mobility Predictions", In *INFOCOM*, March 2004.

[39] D. S. Lee and Y. H. Hsueh, "Bandwidth-Reservation Scheme Based on Road Information for Next-Generation Cellular Networks", *IEEE Transactions on Vehicular Technology*, Vol. 53, No. 1, January 2004.

[40] N. Samaan and A. Karmouch, "A Mobility Prediction Architecture Based on Contextual Knowledge and Spatial Conceptual Maps", In *IEEE Transactions on Mobile Computing*, Vol. 4, No.6, November 2005.

[41] J. Scourias and T. Kunz. "Activity-based Mobility Modeling: Realistic Evaluation of Location Management Schemes for Cellular Networks", In *IEEE Wireless Communications and Networking Conference (WCNC 1999)*, pages 296 – 300, September 1999.

[42] W. Ma, Y. Fang, and P. Lin, "Mobility Management Strategy Based on User Mobility Patterns in Wireless Networks", *IEEE Transactions on Vehicular Technology*, Vol 56, No. 1, January 2007.

[43] J. Ghosh, M. J. Beal, H. Q. Ngo, and C. Qiao, "On Profiling Mobility and Predicting Locations of Wireless Users", *Proceedings of the 2nd International Workshop on Multi-hop Ad-Hoc Networks, from Theory to Reality*, May 2006.

[44] I. Stepanov, P. J. Marrán and K. Rothermel, "Mobility Modeling of Outdoor Scenarios for MANETs", *Proceedings of the IEEE 38th Annual Simulation Symposium*, ANSS 2005.

[45] J. Lessmann and S. Lutters, "An Integrated Node Behavior Model for Office Scenarios", *Proceedings of the IEEE 41st Annual Simulation Symposium*, 2008.

[46] M. Papadopouli, M. Moudatsos, and M. Karaliopoulos, "Modeling Roaming in Large-scale Wireless Networks using Real Measurements", *IEEE Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks* , 2006.

[47] F. Chinchilla, M. Lindsey, and M. Papadopouli, "Analysis of Wireless Information Locality and Association Patterns in a Campus", *IEEE INFOCOM*, 2004.

[48] M. Kim, D. Kotz, and S. Kim, "Extracting a Mobility Model from Real User Traces", *IEEE INFOCOM*, 2006.

[49] M. Boc, A. Fladenmuller and M. D. de Amorim, "Towards Self-Characterization of User Mobility Patterns", *IEEE 16th IST Mobile and Wireless Communications Summit*, July 2007.

[50] J. Yoon, B. Noble, M. Liu, and M. Kim, "Building Realistic Mobility Models from Coarse-Grained Traces", *MobiSys*, 2006.

[51] A. Adas, "Traffic Models in Broadband Networks", *IEEE Communications Magazine*, July 1997.

[52] Y. Suzuki, "Prediction of Daily Traffic Volumes by Using Autoregressive Models", *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, October 1999.

[53] V. Bermudez, D. M. Rodriguez, C. Molina and K. Basu, "Adaptability Theory Modeling of Time Variant Subscriber Distribution in Cellular Systems", In *IEEE Vehicular Technology Conference VTC 1999*, vol. 3.

[54] V. Borrel, M. Dias de Amorim, and S. Fdida, "A Preferential Attachment Gathering Mobility Model", *IEEE Communications Letters*, 9(10), October 2005.

[55] B. Zhou, K. Xu, and M. Gerla, "Group and Swarm Mobility Models for Ad Hoc Network Scenarios using Virtual Tracks", *MILCOM*, 2004.

[56] S. Chen, W. Wang and G. Ren, "A Hybrid Approach of Traffic Volume Forecasting Based on Wavelet Transform, Neural Network and Markov Model", *IEEE International Conference on Systems, Man and Cybernetics*, 2005.

[57] L. A. Andriantiatsaholiniaina, L. Trajkovic, "Analysis of User Behavior from Billing Records of a CDPD Wireless Network", *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks*, (LCN'02), 2002.

[58] R. Hutchins and E. W. Zegura, "Measurements from a Wireless Campus Network", *Proceedings of the IEEE International Conference on Communications*, ICC 2002, vol. 5.

[59] D. Tang and M. Baker, "Analysis of a Metropolitan-Area Wireless Network", *Kluwer Academic Publishers, Wireless Networks 8*, 2002.

[60] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN", ACM SIGMETRICS Performance Evaluation Review, 30 (1), June 2002.

[61] S. Almeida, J. Queijo and L. Correia, "Spatial and Temporal Traffic Distribution Models for GSM", *IEEE VTS 50th Vehicular Technology Conference*, 19-22 September 1999.

[62] X. Ning, Z. Li and Y. Zhang, "A Practical Research on Visualized Spatial Analysis of Traffic Volume Data", *IEEE Proceedings of Intelligent Transportation Systems*, vol. 1, 2003.

[63] K. Tutschku and P. Tran-Gia, "Spatial Traffic Estimation and Characterization for Mobile Communication Network Design", *IEEE Journal on Selected Areas in Communications*, 16(5), June 1998.

[64] F. Ashtiani and J. A. Salehi, "Mobility Modeling and Analytical Solution for Spatial Traffic Distribution in Wireless Multimedia Networks", *IEEE Journal on Selected Areas in Communications*, 21(10), December 2003.

[65] G. Hampel, M. J. Flanagan, L. M. Drabeck, J. Srinivasan, P. A. Polakos, and G. Rittenhouse, "Capacity Estimation for Growth Planning of Cellular Networks in the Presence of Temporal and Spatial Traffic Fluctuations", In *IEEE 61st Vehicular Technology Conference VTC 2005*, 2005.

[66] Jong-Kwon Lee and Jennifer C. Hou, "Modeling Steady-State and Transient Behaviors of User Mobility: Formulation, Analysis, and Application", *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc - May 2006.

[67] G. Le Grand and E. Horlait, "A Predictive End-to-End QoS Scheme in a Mobile Environment", *Proceedings of the Sixth IEEE Symposium on Computers and Communications*, July 2001.

[68] B. Benmammar and F. Krief "Resource Management for End-to-End QoS in a Mobile Environment", *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, June 2006.

[69] J. François and G. Leduc, "Entropy-Based Knowledge Spreading and Application to Mobility Prediction", *Proceedings of the 2005 ACM Conference on Emerging Network Experiment and Technology*, October 2005.

[70] E. Cinlar, *Introduction to Stochastic Processes*, Prentice Hall, 1975.

[71] E. Chlebus and W. Ludwin, "Is Handoff Traffic Really Poissonian?" *IEEE International Conference on Universal Personal Communications*, November 1995.

[72] W. L. Winston, *Operations Research : Applications and Algorithms*, Duxbury Press, 2003.

[73] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*, Dover Publications Inc., 1981.

[74] A. Horváth and M. Telek, "PhFit: A General Phase-Type Fitting Tool", *12th International Conference on Modelling Tools and Techniques for Computer and Communication System Performance Evaluation*, vol. 2324, 2002.

[75] A. Panchenko and A. Thümmler, "Efficient Phase-Type Fitting with Aggregated Traffic Traces", *Performance Evaluation*, vol. 62, 2007.

[76] R. A. Howard, *Dynamic Probabilistic Systems: Volume II - Semi Markov and Decision Processes*, Wiley, 1971.

[77] M. S. Sricharan and V. Vaidchi, "A Pragmatic Analysis of User Mobility Patterns in Macrocellular Wireless Networks", *IEEE Symposium World of Wireless, Mobile and Multimedia Networks*, June 2007.

[78] M. S. Sricharan and V. Vaidehi, "Mobility Patterns in Macrocellular Wireless Networks", *IEEE International Conference on Signal Processing, Communications and Networking*, ICSCN February 2007.

[79] E. Halepovic and C. Williamson, "Characterizing and Modeling User Mobility in a Cellular Data Network", *ACM Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks*, PE-WASUN 2005.

[80] J. Almhana, Z. Liu, V. Choulakian, and R. McGorman, "A Mobile Terminal Location Tracking Model for Personal Communication Systems", *Proceedings of the IEEE Conference on Local Computer Networks*, LCN 2005.

[81] http://www.ee.ucl.ac.uk/∼mflanaga/java/

[82] http://www.mathworks.com/products/matlab/

[83] http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus

[84] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer, "Comparison of Different Methods for Next Location Prediction", *(Euro-Par) - Lecture Notes on Computer Science*, Springer 2006.

[85] D. Katsaros and Y. Manolopoulos, "Prediction in Wireless Networks by Markov Chains", *IEEE Wireless Communications*, April 2009.

[86] H. B. Kim. "An Adaptive Bandwidth Reservation Scheme for Multimedia Mobile Cellular Networks", In *IEEE International Conference on Communications (ICC) 2005*, May 2005.

[87] M. Shankar, M. De Miguel, and J.W.S. Liu, "An End-to-End QoS Management Architecture", *Proceedings of the Fifth IEEE Real-Time Technology and Applications Symposium*, June 1999.

[88] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, "Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications", *Wiley-Interscience Publishing*, 2006.

[89] http://netserver.ics.forth.gr/datatraces/

[90] T. Camp, J. Boleng and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research", *Wireless Communications and Mobile Computing*, Volume 2, 2002.

[91] J. Yoon, M. Liu and B. Noble, "Random Waypoint Considered Harmful", *IEEE INFOCOM*, April 2003.

[92] F. Theoleyre, R. Tout and F. Valois, "New Metrics to Evaluate Mobility Models Properties", *IEEE International Symposium on Wireless Pervasive Computing*, February 2007.

[93] C. Bettstetter, "Smooth is Better than Sharp: A Random Mobility Model for Simulation of Wireless Networks", *Proceedings of ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, July 2001.

[94] M. Zhao and W. Wang, "A Novel Semi-Markov Smooth Mobility Model for Mobile Ad Hoc Networks", *IEEE GLOBECOM*, 2006.

[95] J. Yoon, M. Liu and B. Noble, "Sound Mobility Models", *MOBICOM*, September 2003.

[96] J. Le Boudec and M. Vojnovic, "Perfect Simulation and Stationarity of a Class of Mobility Models", *INFOCOM*, 2005.

[97] W.-J. Hsu, T. Spyropoulos, K. Psounis and A. Helmy, "Modeling Spatial and Temporal Dependencies of User Mobility in Wireless Mobile Networks", *IEEE/ACM Trans. Networking*, Oct. 2009, vol. 17, No. 5.

[98] Y. Lin and V. Mak, "Eliminating the Boundary Effects of a Large-Scale Personal Communication Service Network Simulation", *ACM Transactions on Modeling and Computer Simulation*, Vol. 4, No. 2, April 1994.

[99] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, December 1997.

[100] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. C. Diot, "Packet-Level Traffic Measurements from the Sprint IP Backbone", *IEEE Network*, December 2003.

[101] H. Abu-Ghazaleh and A. S. Alfa, "Mobility Prediction and Spatial-Temporal Traffic Estimation in Wireless Networks", In *IEEE 67th Vehicular Technology Conference VTC2008-Spring*, May 2008.

[102] H. Abu-Ghazaleh and A. S. Alfa, "Application of Mobility Prediction in Wireless Networks Using Markov Renewal Theory", In *IEEE Transactions of Vehicular Technology*, (Accepted for Publication).

LIST OF ACRONYMS

3G 3rd Generation of mobile networking technology

AP Access Point

CDMA Code-Division Multiple Access

CDPD Cellular Digital Packet Data

FDMA Frequency-Division Multiple Access

GSM Global System for Mobile communication

MRP Markov Renewal Process

QoS Quality-of-Service

TDMA Time-Division Multiple Access

WLAN Wireless Local Area Network